Differentiation across the *Podisma pedestris* hybrid zone inferred from high-throughput sequencing data

Hannes Becher

School of Biological and Chemical Sciences,
Queen Mary University of London

Submitted: December 19, 2017

Submitted in partial fulfilment
of the requirements of the Degree of
Doctor of Philosophy

Statement of originality

I, Hannes Becher, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

Signature:

Date: December 19, 2017

Details of collaboration and publications: The work towards Chapter 4 was carried out under the supervision of Dr Konrad Lohse who introduced me to Mathematica. Four perl scripts for the paralogy analysis were provided with adjustments by Beate Nürnberger. These were used originally to analyse the data of Nürnberger et al. (2016). The results presented here, the analyses, and conclusions drawn are my own.

Abstract

Hybrid zones are regions where genetically differentiated forms come together and exchange genes through hybrid offspring. The study of characters gradually changing across such zones (clines) can give insight into evolutionary processes, providing exceptionally sensitive estimates of the intensity of selection, and allowing the detection of loci that might be involved in reproductive isolation and speciation. The Alpine grasshopper *Podisma pedestris* has a hybrid zone in Southern France where two populations meet. They differ in their sex chromosome system, and strong selection against hybrids is observed. These distinct populations likely have split and re-joined several times during the Quaternary glacial cycles. A model explaining the selection observed against hybrids postulates hundreds of loci of small effect spread over two differentiated genomes meeting in secondary contact. Yet, over 50 years of study to-date non have been discovered. However, so far the study of *P. pedestris* has not made use of high-throughput sequencing data which provides an unprecedented resolution of molecular markers.

I am aiming to close the gap with this thesis. I assemble the grasshopper's mitochondrial genome sequence and infer what proportion of its genome is made up by mitochondrial inserts (Numts). Using transcriptome data from two individuals, I then go on to fit demographic models, finding the populations split approximately 400 000 years ago and that the current-day population sizes are considerably smaller than the ancestral one. The final data chapter explores the genetic architecture of the hybrid zone using data from a targeted sequence capture of hundreds of loci covering some 10 000 polymorphic sites. Only two loci under selection are identified, which is surprising given the power of the analysis. Both loci are located on the X chromosome and are subject to weak selection (0.3 % and 0.03 %). This shows the power of hybrid zone analysis to infer targets of selection. The results are discussed in light of a theoretical chapter on the 'inexorable spread' phenomenon and lead to the proposal for further research into the causes of the reproductive isolation observed between the grasshopper populations.

Acknowledgements

This work marks the end of my PhD studies and the end of an important period in my life. On a different scale – my journey into evolutionary genetics – it marks a mere checkpoint on an exciting way along which I am looking forward to move on.

My journey on this way began when I met Professor Richard Nichols who later became my PhD supervisor (in German more poetically "Doktorvater"). Richard deserves thanks for many reasons. He was convinced that I had to do a PhD in evolutionary genetics, he allowed me to work on a fascinating project, he supported me scientifically and personally.

But there was a time before Richard. The first ones to receive me and make me welcome in England were Professor Andrew Leitch and his group with whom I spent many happy hours working, chatting, and travelling. In particular, I want to thank Lu Ma, Stephen Dodsworth, Wencai Wang, Maïté Guignard, and Sara Seco. I owe thanks to Ilia Leitch at Kew Gardens who introduced me to flow cytometry, a means of accurately determining genome sizes. But not only had the Leitches a scientific influence, they are also skilled grasshopper catchers who helped on the 2015 field trip. Three generations of project students helped, too, and I am particularly grateful to Martin Garlovsky, who later became my London house mate and a connection to Sheffield.

The first half of my London time, I spent in the house of Nat Khalawan who is an extraordinary man never ceasing to surprise with reflections on science, politics (both domestic and global), cooking, and the jet stream. He was exceedingly kind and I owe him great thanks for his hospitality.

I met Nat through Raj Joseph, lab technician and famous cricketer with his big, orange-gloved, helping hands. I also wish to thank Monika Strübig, Phil Howard, and Paul Fletcher for technical support both in the lab and the field.

During my time in Edinburgh I was supervised by Konrad Lohse whom I want to thank for his advice and time. Beate Nürnberger helped me with the technicalities of paralogue identification. Edinburgh's Institute of Evolutionary Biology is a hugely inspiring environment and I appreciate greatly the Genetics and Classic Paper Journal Clubs as well as the Evolutionary Lab Group Meeting.

I was welcomed warmly by the crew of the writing-up office 102 whom I wish to thank, too. I also want to thank my short-term house mate Ben Jackson whose presence was very welcome and whose cooking I enjoyed greatly.

Funding for my journey was provided by the EU's Leonardo-da-Vinci programme who supported the first six months of my stay in Andrew's lab, by Queen Mary's School of Biological and Chemical Sciences who awarded me a PhD studentship, and by the Genetics Society who gave me a training grant for my stay at the University of Edinburgh. I wish to thank all of them sincerely.

All this would not have been possible without the support of my parents who have my eternal gratitude. Finally I wish to thank my partner Kim Prior who encouraged me to make a visit to Edinburgh for which I am thankful for several obvious reasons. While finishing her own PhD she made sure I had a great time providing both critical review and welcome distraction when needed.

Edinburgh December 2017

Hannes Becher

Contents

List of Figures

List of Tables

1 General introduction

1.1 The study system

Podisma pedestris (Linnaeus, 1758) is a species of Acridid grasshoppers. It shows a considerable sexual dimorphism (see Figure 1). Males are 17–19 mm long with abdomens striped black and yellow. Their femurs are red below and the tibias are blue. Females are 24–30 mm in length and more camouflaged in tones ranging from brown to green. Both sexes have a white stripe running along the dorsal side of their abdomen, and both sexes are micropterous and flightless.

Figure 1: **Sexual dimorphism.** A male (top) and a female (bottom) individual of *Podisma pedestris*.

The species' distribution covers wide parts of Northern Siberia. In Europe, it occurs in Fennoscandia, further south it is only found in mountain ranges. In the Alps and Pyrenees it usually occurs at altitudes of 1400–2500 m (see Nichols & Hewitt (1986) and references therein) where the species is univoltine (i.e., there is one generation per year). Generations do not overlap as only eggs overwinter. Because it is flightless, *P. pedestris* does not disperse very far. For the parent-offspring-dispersal distance, Barton & Hewitt (1982) determined a standard deviation of approximately 20 m (in one dimension).

Over most of its range, *P. pedestris* has an X0 sex-chromosome system with a karyotype comprising eleven pairs of autosomes and X chromosomes (one in males, two in females). All of these chromosomes are acrocentric. In the Southern French Alps, a population with a different karyotype was found, carrying a neo-X/neo-Y sex-chromosome system created by an X-to-autosome fusion. This fusion created a big, metacentric neo-X chromosome, leaving one single autosome which is called a neo-Y chromosome, because where the neo-X is fixed it can only occur in males (the females having two X-chromosomes). Thus, the karyotype of the fused population consists of ten pairs of autosomes and two neo-X chromosomes in females or one neo-X and one neo-Y chromosome in males (John & Hewitt, 1970), see also Fig. 2 on page 18.

1.2 What is known?

Almost fifty years have passed since the chromosomal polymorphism was discovered and much has been learned about *P. pedestris*. (Hewitt & John, 1972) showed that recombination between the neo-sex chromosomes is significantly reduced and that recombination is displaced away from their centromeres. Hewitt (1975) discovered that the chromosomal populations meet and that there are mixed populations forming a narrow hybrid zone running from Tende in the East near the French-Italian border to Seyne-les-Alpes about 100 km west. While this zone was analysed, it inspired theoretical work on hybrid zones and clines under neutrality and selection (Barton, 1979b,a, 1983). It was discovered that there is strong selection against hybrids (Barton, 1980) and an explanatory model was conceived (Barton & Hewitt, 1981b). The dispersal of *P. pedestris* was estimated to have a standard deviation of 20 m per generation *in one spatial dimension* (Barton & Hewitt, 1982). The standard deviation is given here because the mean dispersal distance is 0. The chromosomal cline was precisely mapped geographically and a model for the colonisation biology of *P. pedestris* was developed (Nichols & Hewitt, 1986). The species' topographical and ecological preferences were analysed and no differences were found between the populations (Nichols & Hewitt, 1988). Also, there is no indication of assortative mating in *P. pedestris*, but there is some evidence for postmating

prezygotic isolation (Hewitt et al., 1987), prompting the question of whether this has evolved as a result of secondary contact.

While there is a stable cline of more or less constant width and strong selection against hybrids was observed independently in three studies (summarised in Nichols & Hewitt (1988)), no molecular differences were found between the chromosomal populations in the era before the advent of affordable high-throughput sequencing. Allozyme studies did not find consistent differences (Halliday et al., 1983, 1984) between populations, and studies on rDNA found exchange of nucleotide sequences (Keller et al., 2008) and foci identified via flourescent-*in situ*-hibridisation (Veltsos et al., 2009) across the hybrid zone.

1.3 Clines and hybrid zones

The word cline was suggested by Huxley (1938) to describe

> a gradation of measurable characters [...] Prefixes can be used to denote different types, for example, ecocline, genocline (gradient in genes), geocline (geographical cline), chronocline (paleontological trend), etc.

Today the word is mainly used for characters gradually changing in space. Hybrid zones, the interfaces between diverged populations where intermediate progeny can be found, are natural places to look for clines. Called 'natural laboratories', see for instance Hewitt (1988), hybrid zones and clines therein allow us to study evolution in action. The analysis of cline width and shape allows us to infer the strength of selection acting and the number of loci involved (Barton & Gale, 1993). While these properties could be analysed, in theory, in crossing experiments, such experiments are often time-consuming or impossible to carry out with large-enough numbers of individuals, in particular for many non-model organisms. For instance, given the width of the chromosomal cline in *P. pedestris* and estimates of its dispersal, it can be calculated that selection of the order of 1 % is acting against chromosomal heterozygotes. It would be extraordinarily hard to pick up such a weak selection coefficient in a crossing

experiment which would require to screen more than 100 000 offspring.[1] The hybrid zone approach becomes even more powerful if there is the possibility to comparatively analyse multiple transects. While neutral genetic variation may by chance show clinal patterns resembling selection, such trends are most unlikely to be observed in replicate transects. High-throughput sequencing enables the generation of unprecedented amounts of data. Applied to a hybrid zone setting, it is perhaps the most powerful way to identify selection acting in genomes.

1.4 Aims of this thesis

As outlined above, much has been found out about the *P. pedestris* hybrid zone in the last half-century. This thesis describes the first study of the zone to make use of high-throughput sequencing data, allowing the re-evaluation of earlier predictions and the generation of new tools for future studies. In particular, this thesis aims to:

1. Investigate whether the distribution of the X0/neo-XY sex chromosomes agrees with ranges of the populations which re-colonised the Alps after the last glacial maximum.

2. Assemble the mitochondrial genome of *P. pedestris* and to analyse the genomic content of Numts both of which will be useful for future studies.

3. Investigate the age of the split between populations meeting in the hybrid zone.

4. Analyse the genetic architecture of the hybrid zone and to assess genomic differentiation across the hybrid zone.

[1] Assuming a biallelic locus with selection against heterozygotes of 1 %, a crossing experiment needs to include enough offspring to decide whether the genotype frequencies observed are significantly different from Hardy-Weinberg proportions (the expectation without selection). The question is how many observations are needed in order for the 95 % quantile of a binomial distribution (one-tailed confidence interval) to be $< 2pq$. This number depends on the genotype frequencies. In the simplest case, $p = q = 0.5$, it is approximately 105 000. The more different the allele frequencies, the smaller becomes the difference between HW frequencies and the expectation with selection and, thus, more offspring will need to be genotyped.

1.5 Structure of the thesis

This thesis is written in paper style and the four data chapters following are planned for publication:

Chapter 2 explores the potential role of 'inexorable spread' (Veltsos et al., 2008) in *P. pedestris*. Inexorable spread is a term coined to describe phenomena identified in simulations that can explain the replacement of an ancestral sex chromosome system by a neo-XY system driven by sexually antagonistic selection.

The chapter describes a simulation-based study. The code was written from scratch in the julia language. Briefly, a grid of demes with adjustable sizes is considered, with stepping-stone dispersal. The chapter also makes use of elevation data from the actual range of *P. pedestris*, which was used to infer the position of watersheds.

Chapter 3 makes use of genome skimming data generated by Illumina's short paired-end NextSeq and PacBio's long read RSII technologies in order to assemble the *P. pedestris* mitochondrial sequence, and to investigate the genomic proportion of nuclear inserts of mitochondrial DNA (Numts). Data generated by both sequencing methods are differentially enriched for mitochondrial sequences and Numts. The importance of Numts is discussed with respect to the species' large genome size.

Chapter 4 aims to shed light on the age of the *P. pedestris* populations which meet in the hybrid zone. It is based on transcriptome sequencing data. Polymorphism data from two diploid individuals are summarised in blocks of four-fold degenerative sites (the block-wise site frequency spectrum) and demographic models under the coalescent are fitted to the data using a Maximum Likelihood-based method developed by Dr Konrad Lohse.

Chapter 5 explores the genomic architecture of the hybrid zone and of the hybrid inviability observed. It is based on targeted capture sequencing data

from hundreds of loci across the *P. pedestris* genome. Clines are fitted to the data and their width and concordance between transects is analysed.

Chapter 6, the general discussion, summarises the results and sets out further directions.

The appendices A, B, and C correspond to the data chapters 2, 3, and 5, **Appendix D** contains my CV, and **Appendix E** contains my paper on a virus-derived genomic repeat in *Fritiallaria imperialis*, the crown imperial fritillary. Work carried out towards this study influenced my PhD, in particular the analysis of genomic repeats with the RepeatExplorer package which gave raise to Chapter 3.

2 *Podisma pedestris* and the inexorable spread

Inexorable spread is a hypothetical phenomenon explaining the spread of a neo-sex chromosome system. It was suggested based on simulations of two hybridising populations differing by an X-to-autosome fusion. Such a fusion would create a larger neo-X and a neo-Y chromosome (homologous to the autosome fused to the X). If sexually antagonistic selection is acting on the neo-Y chromosome, then the neo-sex chromosome system may be spreading, because some crosses produce females carrying Y chromosomes, who will have reduced fitness. These females must carry at least one copy of the un-fused X in order to accommodate a Y, hence this system produces indirect selection in favour of the large fused neo-X. Once these chromosomes are present at high frequency, neo-Ys can follow. This chapter explores whether the inexorable spread may be of importance to *P. pedestris* and what its implications would be on geographic and genomic level.

2.1 Indroduction

Hybrid zones are fascinating study sites where reproductively isolated populations meet and exchange genes. The nature of this gene flow can give an insight into the origin and age of a hybrid zone, the isolating mechanism acting, and, more generally, the evolution of differentiation among species (Barton & Hewitt, 1985). Hybrid zones may arise *in situ* if genetic incompatibilities evolve and become fixed in two subpopulations or they may originate though secondary contact of diverged forms. Looking at a hybrid zone it can be difficult to tell which path has lead to their establishment. However, by considering the Quaternary climatic changes whose cycle of glaciation and milder interglacials made species oscillate between high-altitude/northerly ranges and lower-altitude/southerly refugia, it has been argued that many hybrid zones run in locations where secondary contact would have taken place (Hewitt, 2000) and, more poetically, Hewitt (2001).

The exact location of a hybrid zone depends not only on historic species movements, but also on the isolating mechanism acting. If populations have different ecological preferences, hybrid zones will occur in ecotones. If, however, reproductive isolation has purely genetic causes, that is, hybrid genotypes are intrinsically less fit than parentals, then the hybrid zone position depends on the interplay of parental dispersal and selection against hybrids. Such zones are called tension zones and they are predicted to move and get trapped in areas of low population density (Barton, 1979b).

Hybrid zones are often seen as spatially stable. But this classification is clearly subject to the time scale applied. Slow movement may not be easy to detect without temporally replicated data while swift movement should be more obvious but will last only until one population has completely replace the other. Buggs (2007) compiled a list of moving hybrid zones driven by climate change, human-induced changes in ecology, and genetic causes.

A novel concept for hybrid zone movement was put forward by Veltsos et al. (2008) inspired by the grasshopper *Podisma pedestris*. The species is known for a particularly well-studied hybrid zone in the French Alps. There are two populations of the grasshopper which differ in their sex chromosome system. While one has an X0 system (there is no Y, only one X in males, and two X chromosomes in females), the other has a neo-X/neo-Y system generated by an X-to-autosome fusion (John & Hewitt, 1970). The sex-chromosomal karyotypes of both populations and their possible hybrids are shown in Figure 2. Occurring predominantly above 1500 m in the southern Alps, the populations meet and hybridise in a narrow zone where strong selection is observed against hybrid genotypes. The selection has been assessed by crosses in the lab and by developmental analysis of egg pods collected from the field (Barton, 1980; Barton & Hewitt, 1981b). The chromosomal polymorphism itself seems not to have a major contribution to the hybrid inviability observed, since karyotyping of embryos produced by hybrid females suggests that non-disjunction is rare. Furthermore, the fusion cline is wider than expected were the inviability observed in hybrids (Barton, 1980) due to strong selection on the fusion. None of the previous studies on *P. pedestris* found consistent differences between the populations. The search for differences has encompassed vegetation, geological,

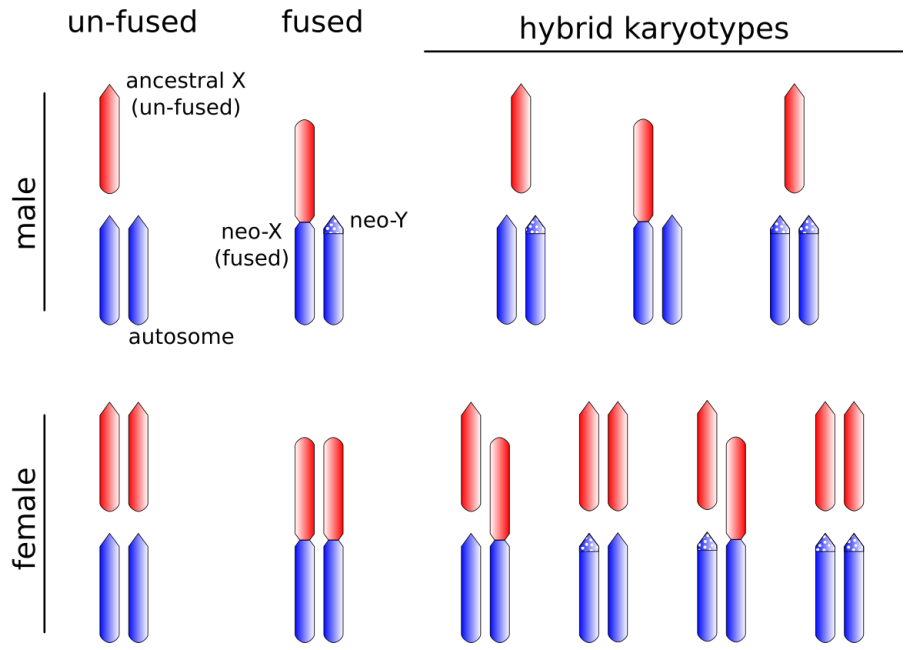

Figure 2: **Sex-chromosomal karyotypes found in *P. pedestris* and its hybrids.** Male karyotypes are shown in the top half, females bottom. The ancestral configuration X0 (un-fused) is shown on the left, followed by the derived neo-XY (fused) configuration where a blue autosome and the red X chromosome are fused. An area of reduced recombination on the neo-Y is indicated by dots. Possible hybrid karyotypes are shown on the right. Note it is possible for the neo-Y to be passed on into hybrid females.

and topographic preferences (Nichols & Hewitt, 1988) as well as genetic differentiation of allozymes (Halliday et al., 1983, 1984). There is, however, evidence for postmating prezygotic isolation (Hewitt et al., 1987, 1989), although crosses in a congeneric species show that females can reproduce parthenogenically when paired with an inappropriate male, a process which could be misconstrued as an excess of homo-karyotypic fertilisations (Warchałowska-Śliwa et al., 2008).

During the Quaternary glaciation cycles *P. pedestris* would have repeatedly left its Alpine range (which would have been covered by ice during the glacial maxima) and would have returned during periods of more suitable climate. These re-colonisations likely happened along different Alpine valleys (river catchment areas) providing much opportunity of population differentiation

and the evolution of population structure due to effects such as long distance dispersal (Nichols & Hewitt, 1994), the embolism effect (Bialozyt et al., 2006), allele surfing (Klopfstein et al., 2006), and bottlenecks (additionally to drift during times of allopatry). Indeed, Veltsos et al. (2009) showed that the genomic distribution of rDNA in *P. pedestris* agrees with this hypothesis. Different valleys show characteristic rDNA distributions. The areas of first contact between different arms of re-colonisating grasshoppers would have been cols at the ends of river catchment ares.

The concept inspired by *P. pedestris* is the inexorable spread phenomenon: if in a hybrid zone between X0 and neo-X/neo-Y forms there is sexually antagonistic selection acting on the Y, some females will have reduced fitness because they acquire a Y chromosome. This outcome is possible because the sex determining system depends on the number of X-chromosomes and not the presence/absence of the Y. Since females homozygous for the fused neo-X cannot accommodate a Y chromosome, the presence of the Y will give an advantage to the neo-X and once the neo-X spreads the neo-Y can follow.

In this chapter, the model of (Veltsos et al., 2008) is re-implemented and extended to consider gene flow in two dimensions and the effect of finite deme sizes. The results are discussed with regards to the distribution of the populations and the expected position of the hybrid zone derived from the position of major watershed in the area.

2.2 Materials and methods

2.2.1 Modelling

The hybrid zone simulations were carried out in julia (version 0.4.5), a high-level programming language which combines python-like syntax and many available libraries with high performance. The hybrid zone was represented as a grid of 120x10 demes each with a set population size. Each grid cell held a vector of genotype frequencies. The eleven genotypes possible are shown in Figure 2. The grid was split in two along the longer axis and each half was filled with genotypes of one population. The grid was then subjected to 10 000 rounds of migration, selection, sex, and (optionally) genetic drift. Migration

was implemented in stepping-stone fashion with demes dispersing 4 % of their genes in each of the four cardinal directions. Selection on the Y chromosome was set to 1 % in males and −1 % in females with alleles acting additively in both sexes. Selection of 1 % against heterozygous females was included optionally to reflect the possible effect of non-disjunction. These parameters represent the default values chosen by Veltsos et al. (2008) (who explored the consequences of a wider range of parameter combinations.) The action of genetic drift was included optionally. When considered, zygotes were sampled according to the population size chosen (40, 86, 186, or 400) from a pool of genotypes.

For each combination of parameters, ten replicate runs were carried out (except for deterministic cases with infinite populations size). The simulations were run for 10 000 generations and the state of the deme grid was saved every 100 generations. The positions of the X and Y clines were computed after 10 000 generations using GLMs with a logit link function. If alleles had become fixed before generati on 10 000, cline positions were determined at generation 7000 and were extrapolated (this seems to somewhat underestimate the cline speeds).

2.2.2 Watersheds and distribution data

Elevation data on a 90 m-grid was retrieved from the Consortium for Spatial Information (CGIAR-CSI) at `http://srtm.csi.cgiar.org/SELECTION/listImages.asp`. The data in ASCII format were trimmed to the coordinates of the field trip area (44.5–44.0° N, 5.8–7.0° E) and were imported into Mathematica where the position of the watershed was determined using the built-in function `WatershedComponents` specifying starting points in four river valleys. The code is shown in Appendix A. The elevation data was then visualised in R, and overlayed with a contour plot and the position of the watersheds determined in Mathematica. Coordinates of sampling sites of the field trips in 2014–2016 and the associated karyotyping information was plotted on top of the map created (shown in Fig. 5).

2.3 Results

2.3.1 Speed of cline movement

The movement speed of the Y-cline is shown in Fig. 3. The movement speed for the deterministic case without directional gene flow (33 and 18 cm/generation) are lower than the values reported by Veltsos et al. (2008) by a factor of $\sqrt{2}$. This is expected because here dispersal was simulated in two dimensions rather than in one, but using the same dispersal parameters. A linear model (Tab. 1) shows how deme size (categorical), directional gene flow (categorical), and selection against heterozygotes (continuous) affect the position of the Y-cline after 10 000 generations. Small deme size does not only cause higher variance in movement speed (see wider boxes in Fig. 3), but also slows down the cline movement. Increasing deme size speeds up the spread of the cline. In the deterministic case (infinite deme size) the cline moves more than 400 m (4 cm/generation) further than with $N = 40$ (the smallest value investigated). The presence of selection against female heterozygotes slows down the cline movement by 13 cm/generation (1.3 km over 10 000 years), and directional gene flow of the order of 1 % slows down cline movement by 4 cm/generation.

2.3.2 Patchiness

While deterministic models of the inexorable spread can be summarised well by the position and steepness of the chromosomal clines, models with finite-size demes are subject to drift so that allele frequencies can vary considerably. In particular with low deme sizes, patches of unexpected frequencies of one chromosome can be found far from the cline centre and uncoupled from the other chromosome. Fig. 4 shows such a situation where a patch of high Y frequency emerges 1000 m ahead of the Y cline. This patch has a low frequency of fused X chromosomes and it persists for several hundreds of generations.

2.3.3 Watersheds

Fig. 5 shows the positions and kayotyping results for the populations sampled. The areas' watersheds are indicated by fat black lines. XY (blue) and X0 (red)

Table 1: **Factors influencing cline speed.** The effect of population size (N, categories: 40, 86, 186, 400, and deterministic), selection against females heterozygous for the X (SH, categorical 0 or 0.01), and directional gene flow (DGF, continuous). Estimates are given in metres per 10 000 generations. See main text for detailed explanation.

```
Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept)   2830.97      34.46   82.160  < 2e-16 ***
N086           227.12      36.74    6.183 1.34e-09 ***
N186           358.84      36.74    9.768  < 2e-16 ***
N400           392.14      36.74   10.675  < 2e-16 ***
Ndet           417.09      86.00    4.850 1.67e-06 ***
DGF         -41196.59     375.62 -109.677  < 2e-16 ***
SH           -1310.13      25.63  -51.118  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 284 on 484 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.9682,    Adjusted R-squared:  0.9678
F-statistic:  2456 on 6 and 484 DF,  p-value: < 2.2e-16
```

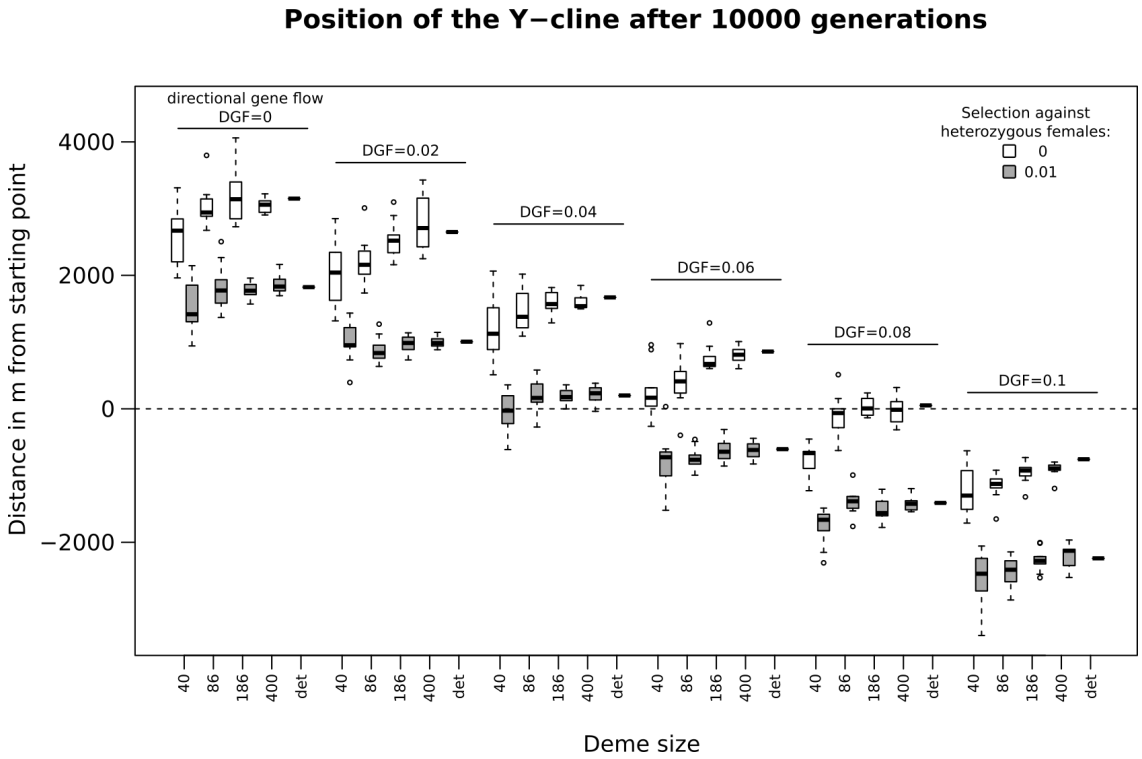

Figure 3: **Movement speeds of the chromosomal clines depend on several factors.** Directional geneflow (DGF), selection against females heterozygous for the X, and small deme size all slow down the spread of the neo-X/Y chromosomes.

populations of grasshoppers are well-separated by the watershed extending from east to west in rows 3 and 4 and turning south in column C. Two transects crossing the hybrid zone (labelled *b* and *c*) cross the watershed too. There are some disagreements: fused grasshoppers occur north of the watershed at *a*. This is not a parapatric site, the populations are separated by a valley. Near *d* the hybrid zone runs across l'Aiguillette, situated approximately 2500 m north of the watershed. At *e*, un-fused grasshoppers occur south of the watershed (Nick Barton, personal communication).

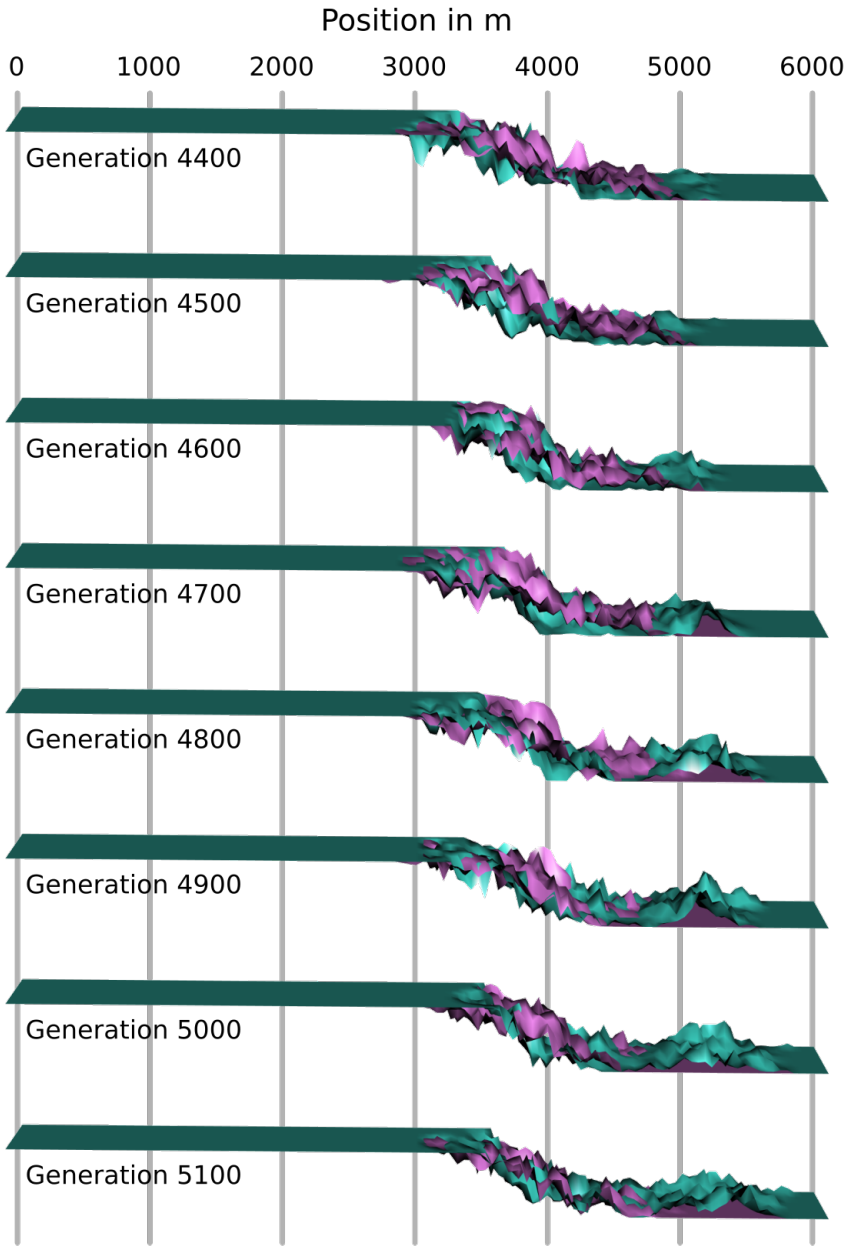

Figure 4: **Slope and cline width of a fitted model do not always characterise a population well.** Here the emergence of an anomaly of high Y frequency 1000 m ahead of the cline is shown, which persists for hundreds of generations (at 5000 m). The chromosomal frequencies are shown in turquoise for the neo-Y and violet for the neo-X.

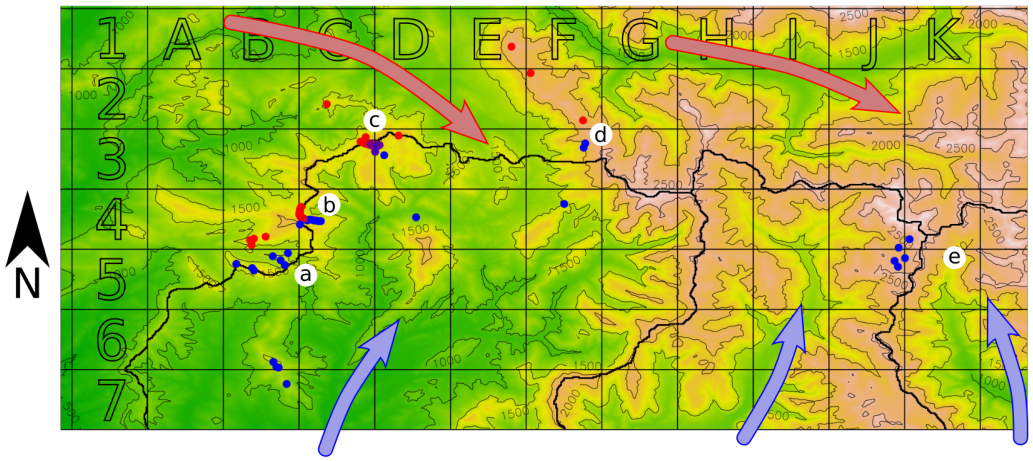

Figure 5: **Distribution of X0 (red) and neo-XY (blue) populations compared to regional watersheds (fat black lines).** Dots represent sampling sites, arrows indicate presumed routes of past colonisation. The grid cells are 5 km. Sites indicated by letters are referred to in the main text.

2.4 Discussion

2.4.1 The inexorable spread – preconditions

The engine of the inexorable spread phenomenon is sexually antagonistic selection acting on a neo-Y chromosome (Veltsos et al., 2008). Is such selection likely to be acting in *P. pedestris*? There are two requirements: a region of reduced or no recombination on the neo-Y and the presence of sexually antagonistic alleles in linkage with this region. There is some evidence the first requirement is met, and studies of other taxa make it seem plausible the second might be, too.

While the chromosomal fusion present in the neo-XY population is believed to go back to one single X-to-autosome fusion event (the fused X is known only from the Southern French Alps and the population's range is continuous),

the neo-Y chromosomes need not have a common ancestor contemporaneous with the X-to-autosome fusion event. In populations where the fusion first became fixed, any remaining autosomes would be trapped in males. This fate would have altered the selection acting on them, and initiated the evolutionary journey leading to the present-day neo-Ys. Alternatively, for instance if the selection on the neo-Y chromosome was instrumental to the spread of the fused sex chromosome system from very early on, both chromosomes may have their most recent common ancestors around the same time. Studies of meiotic chromosome spreads of *P. pedestris* (John & Hewitt, 1970; Hewitt & John, 1972), comparing Y chromosomes to their homologous autosomes in the un-fused population, found that Ys were involved in fewer cross-overs, which were displaced away from the Y's centromere. This was confirmed in samples from multiple subpopulations. The cause of this observation is not clear. Some non-mutually exclusive possibilities are sequence divergence, the presence of an additional rearrangement for instance an inversion, or a spatial effect of the chromosomal fusion (there might simply be no space for crossing-over because of the bulky X-chromosome). Irrespective of the cause, the Y's centromeric area seems to show reduced recombination and thus may be undergoing processes of Y-chromosomal evolution (Charlesworth & Charlesworth, 2000) which could favour the accumulation of more strongly sexually antagonistic alleles, selection for reduced recombination to retain them in the appropriate sex, and genetic degeneration as a consequence of the reduced recombination.

Sexually antagonistic genetic alleles have been shown to segregate in populations of Drosophila (Rice, 1996, 1998). More recent studies support the existence of such diversity in populations of sheep (Robinson et al., 2006), deer (Foerster et al., 2007), and many other taxa (Cox & Calsbeek, 2009). Theory predicts that selection will reduce recombination between sexually antagonistic alleles and sex-linked regions (Rice, 1987; Kirkpatrick & Guerrero, 2014). While there is no evidence for sexually antagonistic selection in *P. pedestris*, its presence would be plausible in light of the species' strong sexual dimorphism.

2.4.2 Spread of a chromosomal fusion and hybrid zone movement

The zone described here has been called a tension zone because there is strong selection against hybrids and the populations hybridising do not show differential ecological preferences (Nichols & Hewitt, 1988). Analyses of crossing experiments have led to the suggestion that the strong selection against hybrids may be due to many loci of small effect which also gives an explanation for why the fusion cline is much wider than selection estimates suggest (Barton & Hewitt, 1981b). This would suggest both sex chromosome systems are restricted to their specific genomic background.

The inexorable spread phenomenon analysed here provides a means for the spread of a chromosomal fusion. This does not necessarily involve the whole genome. If the current chromosome distribution is the result of a spread driven by sexually antagonistic selection, this could mean the chromosomal fusion and the non-recombining portion of the neo-Y would have spread. During this process, recombination might have eroded away loosely linked alleles of the fused X and neo-Y and only a comparatively small amount of genetic material would have spread. In this case, one should not expect to find individuals of different karyotypes to be genetically differentiated across the present zone of contact. Indeed, allozyme studies and analyses of rDNA variants (Keller et al., 2008) did not find strong differentiation across the zone. The lack of differentiation may be in agreement with the hypothesis of chromosomal spread, but it is not particularly strong evidence. Other genetic architectures such as few staggered clines (suggested for *Chorthippus parallelus* (Virdee & Hewitt, 1994)) or strong selection on a single locus linked somewhat to the chromosomal fusion could produce similar selection effects to those observed. These considerations cannot be tested without genetic data, but there is other evidence. It comes from the species' expected colonisation history.

2.4.3 Evidence for the phenomenon

If the inexorable spread phenomenon has acted in the past, then the sex chromosomes' distribution may have changed since the time of their first contact and the fused population would have spread into the un-fused territory. The

parameter space explored in the simulation carried out here suggests advances of 3.5–5 km could be possible. Has this happened? In order to answer this question, a null hypothesis is needed describing the expected position of the populations based on the location of secondary contact and including climate considerations. As the climate gradually warmed after the last glacial maximum, re-colonising populations of grasshoppers presumably moved up Alpine valleys until they met at the 'ends of valleys', the watersheds. Watersheds generally run along ridges and low-elevation points on such ridges are cols. These are the sites where secondary contact is likely to have occurred first. One such area is in cells 3D-F in Fig. 5 where there are three cols below 1500 m (Col du Fanget in the West at 1459 m, the area between Clot des Etables and Le Grand Puy at below 1450 m, and Col de Maure in the East at 1346 m). While *P. pedestris* must have lived there in the past, these areas are too warm now and the grasshoppers have moved onto higher ground to the West and East of these cols.

The present-day distribution (samples indicated as dots in Fig. 5) matches this null hypothesis quite well. The watershed running from 4K to 3C and then turning south-west separates well the fused (blue) and un-fused (red) samples. Two transects crossing the hybrid zone (*b* near Barles and *c* on Les Tomples) lie on the watershed. There are two sites of disagreement where the fused population has advanced north of the watershed (at *a* and *d*). These might be seen as evidence for the hypothesis of inexorable spread occurring. But there are alternative explanations. At site *a* the fused population extends approximately 1 km north of the watershed and the populations do not come into contact, they are separated by a valley. Un-fused grasshoppers may have been cut off from their hinterlands when the valley became too warm and they may have been displaced by the fused population on the high ground. At site *d*, the hybrid zone runs across the mountain l'Aiguillette (2610 m) situated about 2.5 km north of the watershed. The hybrid zone could have moved there after secondary contact first occurred near Col de Maure (cell 3F) where the watershed runs across a shallow area and it is hard to know whether grasshoppers of both populations reached this area at the same time from north and south. Even if contact first occurred exactly on the watershed, the climate improvement during the Holocene forced the hybrid zone to move up onto the Montagne de

la Blanche (the mountain ridge ranging from 1E to 3G), during which neutral processes may have caused the shift of the zone. Another disagreement is the fact that un-fused grasshoppers occur south of the watershed at site *e* (Nick Barton, personal communication). This is the opposite of what is expected under inexorable spread.

Another (weaker) strand of evidence comes from crossing experiments. Between-population crosses show far higher embryo mortality than crosses within populations. Crosses of individuals taken from the centre of the hybrid zone show high embryo mortality, too (Barton, 1980; Barton & Hewitt, 1981b). Both observations have been interpreted as evidence for genomes diverged at many loci coming together, resulting in reduced fitness. Under this model, one would not necessarily expect to find unfit hybrids, associated with the sex-chromosomal cline.

The evidence does not seem to support the presence of inexorable spread in *P. pedestris*. It is clear that the populations' distribution does not exactly agree with the watershed and that other processes were involved in the positioning of the current-day hybrid zone. Inexorable spread might have contributed to this pattern.

2.5 Conclusion

P. pedestris is likely to fulfil the preconditions for inexorable spread to occur. There is evidence for a region of reduced recombination between the neo-sex chromosomes which is likely to catch sexually antagonistic alleles. However, the sex chromosomes' distribution matches quite well the null hypothesis of no change since the first contact. If it has occurred, the neo-sex chromosomes would have moved into a different genomic background which might be detectable by surveying molecular markers from throughout the genome, to obtain evidence for a separation of the sex chromosome transition from clines at other loci. Following this idea, Chapter 5 is about a sequence capture study analysing almost 1000 loci across the *P. pedestris* genome.

3 Mitochondrial sequences or Numts – By-catch differs between sequencing methods

Nuclear inserts derived from mitochondrial DNA (Numts) are fascinating study objects. Being mostly non-functional, and accumulating mutations more slowly than mitochondrial sequence, they act like molecular fossils – preserving information on their ancestral states and carrying further phylogenetic signal encoded in the changes since their insertion into the nuclear genome. These attributes cannot be reliably exploited if Numt DNA sequence is confused with the organellar genome (mtDNA), and the analysis of mtDNA is similarly impeded.

In this Chapter, a method is demonstrated to address this problem of distinguishing Numts from mtDNA, without the need for comprehensive assembly of the nuclear genome or the physical separation of organelles and nuclei. Our approach is to exploit the different biases in alternative sequencing technologies. We find the short-read data yield mainly mtDNA sequences, whereas long read sequencing strongly enriches for Numt sequences. We demonstrate the method using genome-skimming (coverage $\ll$ 1x) data obtained on Illumina's NextSeq and PacBio's RSII platforms from DNA extracted from six grasshoppers (*Podisma pedestris*). The RepeatExplorer package could successfully assemble the mitochondrial genome from the short-read data, despite the presence of Numt reads. We compared short and long-read data to obtain two estimates the genomic proportion of Numts in the *P. pedestris* genome (0.04 % and $\leq$ 0.1 %) a total length equivalent to many hundreds of mitochondrial genomes.

3.1 Introduction

Sequences of mitochondrial DNA have proved indispensable markers for population genetics and phylogenetics for decades (Avise et al., 1987; Ballard & Rand, 2005). More recently, numerous ecological experiments have exploited

the universal animal barcoding marker, COI, which is a mitochondrial gene (Hebert et al., 2003). Because mitochondria are present in high numbers in cells, their DNA tends to amplify easier than its nuclear counterpart rendering mitochondrial DNA useful for research on museum specimens as well as for ancient DNA studies (Mohandesan et al., 2017) and analyses of fecal samples (van der Valk et al., 2017). Thanks to their comparatively small size of approximately 16Kbp and conserved structure (Boore, 1999) animal mitochondrial genomes are easy to assemble (similar to plant plastomes, (Twyford & Ness, 2016)).

However, studies of mitochondrial DNA are considerably impaired by the presence of Numts, nuclear inserts derived from mitochondrial DNA. Evidence of such insertions was first found shortly after mitochondria were discovered to contain their own genetic material (Du Buy & Riley, 1967) and it has since become clear that Numts are present in many species (Bensasson et al., 2001) and often in multiple copies. While they are interesting to study and potentially useful molecular markers, Zhang & Hewitt (1996) argue that Numts have the potential to mislead phylogenetic and population-level studies. It is obvious that knowledge of the genomic content of Numts will be helpful to any study targeting mitochondrial sequences.

Before high-throughput sequencing data became readily available, Numts could be detected, albeit with some difficulty, with hybridisation or PCR-based methods see, for instance, (Gellissen et al., 1983; Bensasson et al., 2000). To physically separate mitochondrial and nuclear DNA, (ultra) centrifugation can be used (Lansman et al., 1981; Garber & Yoder, 1983), but such methods require considerable technical effort. Since the advent of HTS data, there are two main approaches. (1) In well-assembled genomes sequenced at high coverage, Numts can be detected simply by screening the assembly for regions with similarity to mitochondrial DNA (Hazkani-Covo et al., 2010). Where such data exists for multiple individuals, as in humans, polymorphism between Numts can be studied (Dayama et al., 2014). (2) In absence of a well-assembled genome, for instance in genome skimming studies (Straub et al., 2012; Dodsworth, 2015), Numts may be identified by sequence divergence from the actual mitochondrial genome or if reads (or read pairs) match the mitochondrial sequence only partially. Such approaches will fail to detect Numts longer than the sequencing platform's read

length if their sequences are not clearly diverged from mitochondrial ones.

HTS methods are able to produce unprecedented amounts of data. While "second generation sequencing" methods (such as Ion Torrent, Illumina, and 454) generate short reads of some hundreds of bases at high throughput, "third generation sequencing" methods (such as PacBio and Oxford Nanopore) are able to generate reads of several kilobases in length. Because HTS experiments are to be less directed than traditional Sanger sequencing, some fraction of the data generated usually goes beyond the scope of a specific experiment. This fraction, genomic 'by-catch', can be useful since it can be mined for additional incidental information. By-catch can be (and often is) used to assemble genomes of organelles such as mitochondria. The choice of sequencing platform may influence the type of by-catch produced, particularly because of differences in fragmentation and size-selection protocols between second and third generation methods.

Here it is shown that different common sequencing strategies enrich differentially for animal mitochondrial sequences and nuclear insertions of mitochondrial origin (Numts). This became obvious when genome skimming data generated by Illumina's NextSeq and PacBio's RSII platforms from the grasshopper, *Podisma pedestris* were analysed to explore the species' repetitive genome content.

3.2 Materials and methods

3.2.1 Samples and sequencing

Illumina NextSeq Freshly removed hindlegs of *P. pedestris* were snap-frozen and stored at $-79\,^{\circ}\mathrm{C}$. Before DNA extraction, the legs were dipped into boiling water to inactivate DNases. Subsequently the denatured femur muscle were dissected out. DNA was then extracted using a Qiagen Blood and Tissue kit following the manufacturer's instructions. Using a Covaris ultra sonicator the DNA was sheared aiming to achieve a median size of 550 bp. Libraries for sequencing were prepared using an Illumina TruSeq DNA PCR-Free kit. Sequencing was carried out at QMUL's Genome Centre on Illumina's NextSeq Platform using v2 chemistry.

PacBio RSII Freshly removed hindlegs were stored in pure ethanol. DNA was extracted from four samples using a Qiagen Gentra HMW kit resulting in molecules with a lengths of mainly $>$ 48 kbp (TapeStation, Agilent Genomics). Further work was carried out by The University of Liverpool's Centre for Genomic Research. The aimed size for DNA fragmentation was 10 kbp. The four libraries median (non-redundant) read lengths were 3125, 3167 and 2097 bp. Sequencing happened on a PacBio RSII machine using C6 chemistry.

3.2.2 Data cleaning

Two sets of clean NextSeq data were prepared. For the RepeatExplorer analysis, the data were filtered using a custom python script keeping only those read pairs were 90 % of the bases had a phred quality score $>$ 20. A significant amount of the data consisted of adapter dimers. Pairs with reads showing similarity to TruSeq adapters (as identified by BLASTn) were removed.

A second cleaned set of NextSeq data was generated for mapping and variant calling. Here we aimed to remove as many spurious base calls as possible. The first 5 bp of every read were removed and, using Skewer (Jiang et al., 2014), the 3'-ends were trimmed back until each last base's quality was 30 or higher. All PacBio reads (circular consensus sequences, CCS) were used as they were. For the RepeatExplorer analysis, pseudo paired reads of 151 bp with an insert size of 550 bp were cut out of long PacBio reads using custom-made python scripts which depend on the biopython module (`http://biopython.org/`).

3.2.3 RepeatExplorer analyses

RepeatExplorer (RE, `https://galaxy-elixir.cerit-sc.cz/`) is a pipeline to analyse the repetitive genome content from short read genome skimming data. RE performs all-to-all comparisons and generates clusters of similar reads, which often correspond to particular genomic repeats such as transposable elements or satellites. Mitochondrial genomes and rDNA which are present in high copy numbers are usually picked up as well. While NextSeq reads were used as they were, PacBio reads were fragmented *in silico* to obtain pseudo-paired reads. RE was run twice: first only NextSeq data was supplied in

order to assemble the *P. pedestris* mitochondrial genome. In the second run, 100 000 NextSeq read pairs from each of six individuals (N1–N6) and 150 000 PacBio pseudo read pairs from each of three individuals (P1–P3) were analysed jointly to compare the sequencing methods. The pipeline was supplied with a custom annotation database containing the mitochondrial genome sequence of *Schistocerca gregaria* (Genebank NC_013240.1 (Erler et al., 2010)) in the first round and with the *P. pedestris* rDNA and mitochondrial genome in the second run.

3.2.4 Mitochondrial sequence assembly

Eight RE clusters connected by paired reads (244 – 57 – 230 – 205 – 69 – 85 – 102 – 161) showed sequence similarity to *S. gregaria* mitochondrial DNA. Those clusters overlapping consensus sequences were assembled in Geneious R9, forming a reference to which reads of sample N1 were mapped. High coverage and truncated reads at the control region indicated a duplication, which was then added to the reference. Subsequently each of the six NextSeq samples' sequencing data were mapped individually using bowtie2 (Langmead & Salzberg, 2012) with the following command line: `bowtie2 -x <reference> -1 <fw reads> -2 <rw reads> -p 12 -X 20000 --sensitive-local 2> <log file>`. For each of the six alignments 50 % majority rule consensus sequences were created in Geneious. They were annotated automatically using the MITOS WebServer (Version 2 beta (Bernt et al., 2013)).

3.2.5 Mapping and variant detection

In order to detect individual-specific variants, a second round of mapping was performed with NextSeq data. Polymorphisms were called using Geneious' function 'Find Variations/SNPs' with default settings and a minimum allele frequency set to 0.01. The resulting tables were exported to CSV format and were processed interactively in R 3.3.1 (R Core Team 2016).

All PacBio reads were aligned to the mitochondrial assembly using the LAST suite (Kiełbasa et al., 2011). This package performs alignments similar to BLAST, but it takes into consideration how common specific seeds are in the data

Table 2: **Samples for genome skimming.** 'Mito-like' refers to data that might be derived from actual mitochondrial genomes or Numts.

NextSeq	Sample	Origin	High-quality reads	Mito-like (reads)	Max SNP frequency
	N1	Blayeul	25 600 134	0.48 %	15 %
	N2	Blayeul	37 563 526	0.64 %	14 %
	N3	Blayeul	13 034 544	1.17 %	8 %
	N4	Mariaud	35 257 122	0.97 %	9 %
	N5	Mariaud	33 393 734	1.32 %	7 %
	N6	Mariaud	21 813 470	0.44 %	20 %
PacBio	**Sample**	**Origin**	**Data in bp (non-redundant)**	**Mito-like (bp)**	
	PB1	Blayeul	336 152 778	0.04 %	
	PB2	Blayeul	376 785 844	0.03 %	
	PB3	Bournee	248 262 810	0.06 %	

resulting in more accurate estimated of e-values (the probability to obtain a specific alignment in the sequence library supplied.) In brief, the assembly of individual N1 was masked at low complexity regions *(where GC-content was below* 10 %) and was subsequently converted to a LAST database using the scoring scheme NEAR (optimised to find short matches too), preserving all masked regions and additionally masking simple repeats (optimised for high AT-content). Lastal (from the LAST suite) was then run with parameter D set to one thousand times the length of the assembly corresponding to an e-value threshold of 0.001 in BLAST. This means a sequence of the length and composition given is expected to be found by chance one in a thousand trials. Of the resulting hits, only those with alignment lengths above 100 bp were kept. Shorter ones tended to accumulate in un-masked regions of low complexity, not permitting meaningful conclusions about homology.

3.2.6 Specificity estimation and genomic proportion

RepeatExplorer The output of the second RE run carried out was used to compare sequencing technology-specific bias (see Figure 6).

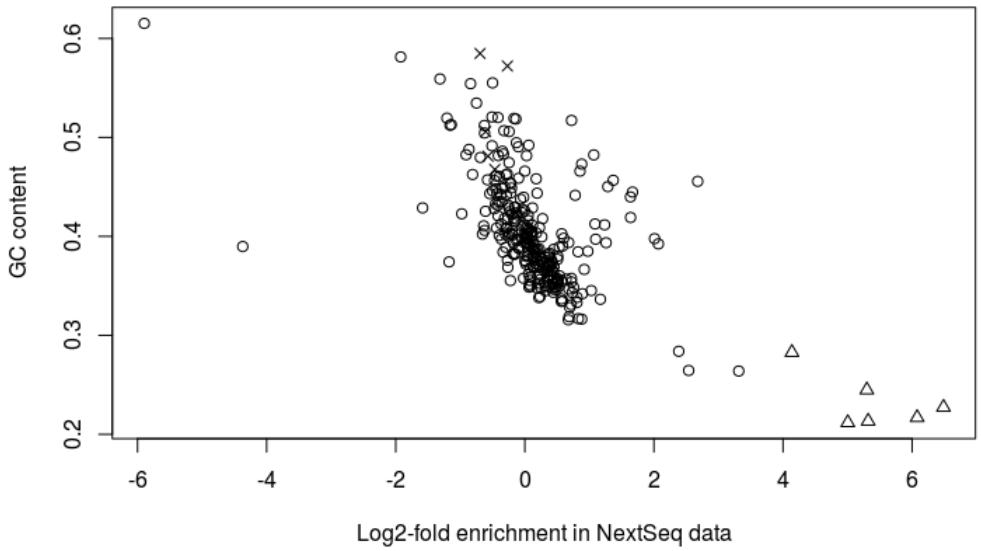

Figure 6: **Log2-fold enrichment of NextSeq reads plotted against GC-content for 300 RepeatExplorer clusters.** There is a general trend for enrichment in NextSeq data for clusters with low GC-content. Mitochondrial-like clusters are indicated using a triangle, rDNA as an X.

NextSeq Sequencing reads mapped to the mitochondrial assembly could either have originated from mitochondrial genomes or from Numts. Assuming each of the individuals contained the same proportion of Numts in the genome, the variation between samples in the proportion of NextSeq reads mapping to the mitochondrial assembly (see Tab. 2) would be attributed to different mitochondrial densities in the extracts. Such differences could have arisen for several reasons, for example the different ages of the samples (some were late larval stages, rather than adults). One estimate of the proportion of mitochondrial reads can be obtained from the assumption that most mitochondrial sequences in any one individual is monomorphic, whereas some of the Numt sequences will be fixed for a different allele. In this case the frequency of this Numt allele will be proportional to the relative contribution of Numts. The maximum of the distribution of allele frequencies (shown as ribbons of jitter in Fig. 7) provides an estimate for the relative contribution of Numts to the data mapping to each

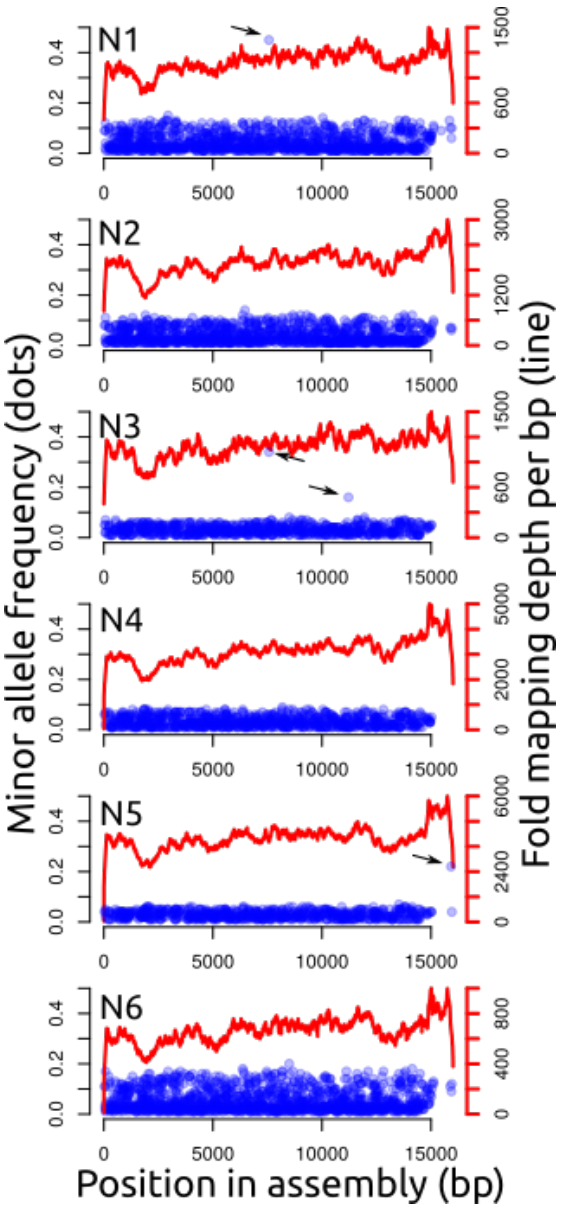

Figure 7: **Polymorphism within assemblies.** For each individual, the minor allele frequencies and positions of SNPs are shown as blue dots. Note, the dots generally form ribbons with widths differing between samples. There are few outliers, marked with arrows, likely indicating heteroplasmy. Individuals N1 and N3 share such a polymorphism at base pair 7567. The red lines indicate per-base pair mapping depths.

mitochondrial assembly.

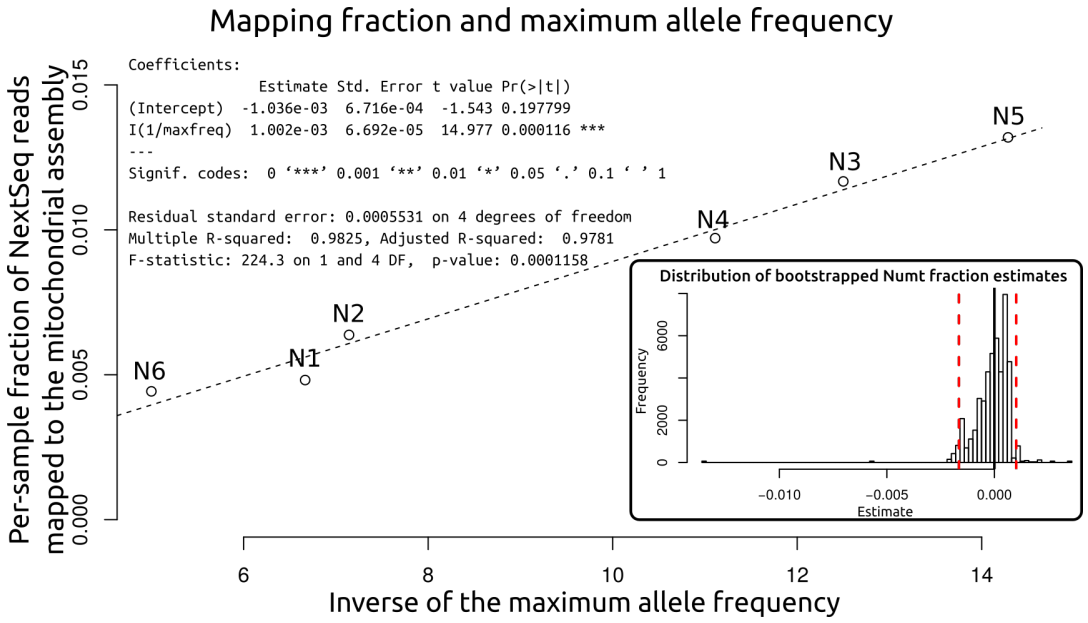

Figure 8: **Relationship between the fraction of NextSeq reads mapped to the mitochondrial assembly and the inverse of the maximum allele frequency of polymorphisms per sample.** The insert shows the bootstrap distribution of the genomic proportion of Numts. Dashed lines indicate the 2.5 and 97.5 percentiles.

This assumption is supported by the very close correlation ($R^2 = 0.98$) between the inverse of the maximum allele frequency and the proportion of all sequencing reads mapping to the mitochondrial sequence (shown in Fig. 8). Extrapolation of the regression line to the point where there are no mitochondrial reads gives a point estimate of the genomic proportion of Numts in *P. pedestris*. An exhaustive bootstrap was performed to assess this estimate's confidence (insert in Fig. 8).

PacBio The specificity of PacBio sequencing was evaluated per read aligned. Reads aligning only partially were considered Numt-derived. Reads matching along $> 95\,\%$ (arbitrarily chosen) of their length were considered full-length matches. For these, alignment error profiles were compared to the reads' phred quality scores. If an alignment contained significantly more mismatches than expected (5 % confidence interval, one-sided, Bonferroni-corrected for 59

alignments), it was considered a Numt sequence. The fraction of potentially mitochondrial reads was so small that the genomic proportion of Numts could be calculated as the ratio of the summed length of Numt sequences and the summed length of all (non-redundant) PacBio data generated.

3.3 Results

3.3.1 RepeatExplorer and repetitive genome content

The RepeatExplorer pipeline (RE, (Novák et al., 2013)) was used to compare abundant sequences in PacBio and NextSeq data. Fig. 6 shows the ratio of reads contributed from each sequencing method to the largest 300 RE clusters plotted against each cluster's GC content. There is a clear effect of GC-content on the bias towards NextSeq data. Generally, the lower a cluster's GC-content, the greater its proportion of NextSeq reads. While the ratio of read origin lies between 0.5 and 2 for most clusters, the mitochondrial clusters show the most extreme enrichment in NextSeq data.

3.3.2 NextSeq: Six mitochondrial genome assemblies for *Podisma pedestris*

For each read cluster identified RE assembles one or more contigs, which can be used to analyse a genomic repeat's sequence and structure. The contigs of RE clusters with similarity to mitochondrial genome sequences were used to generate a draft assembly of the *P. pedestris* mitochondrial genome. Mapping individual-specific reads to the draft assembly, mitochondrial genome sequences of six individuals were generated (for each of the NextSeq samples, PacBio samples did not yield enough mitochondrial sequences for continuous mapping). Each assembly is 16 008 bp long with the control region containing a 383 bp direct repeat. The fraction of sequence reads mapping to the mitochondrial assembly varies between samples, likely indicating differences in cellular content of mitochondria between individuals (Tab. 2). All genes typically found in animals were identified using the MITOS WebServer (v2beta) (Fig. 9A). The gene order is colinear with the ones of other grasshopper mitochondrial genomes, and the sequences align readily (data not shown).

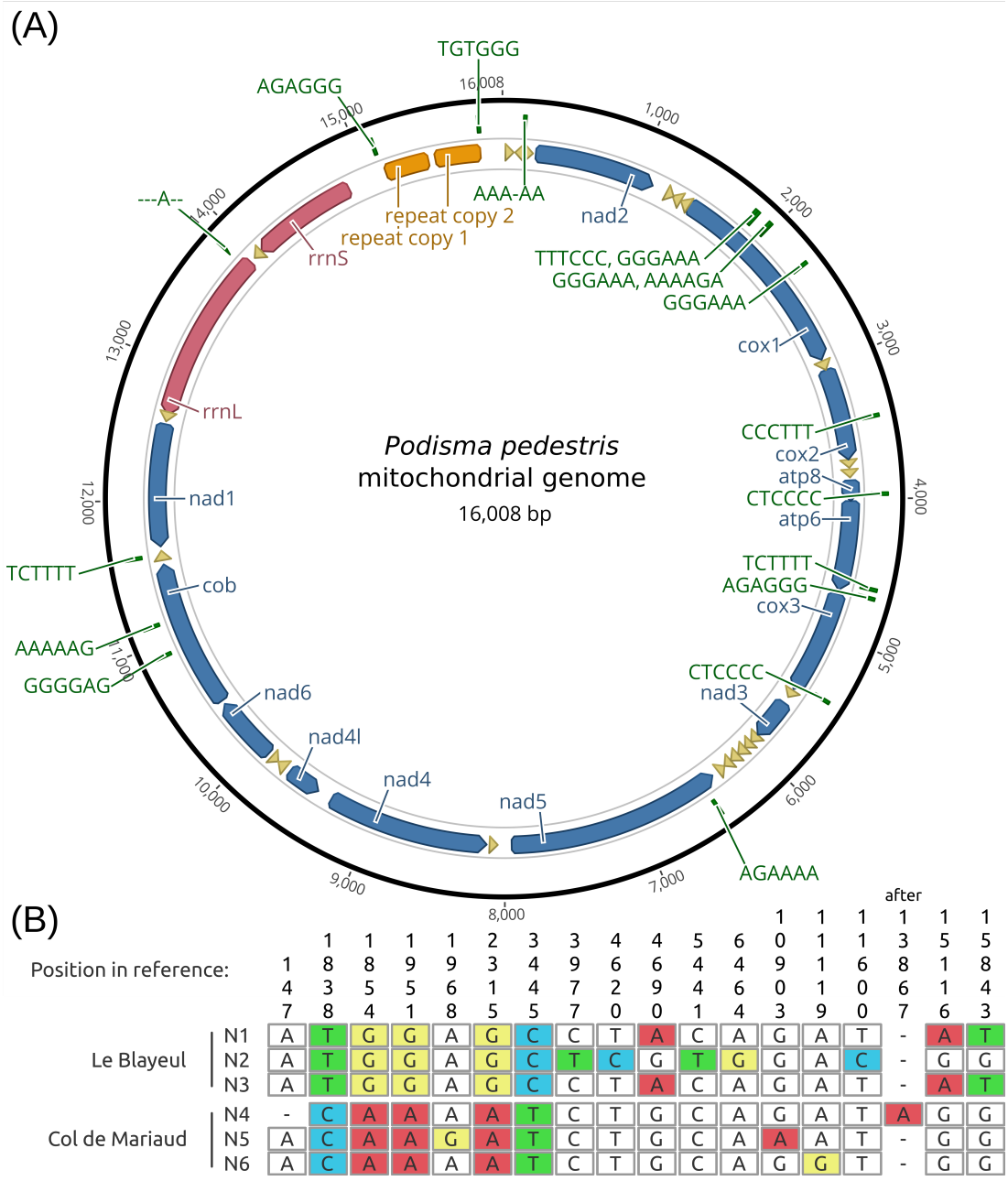

Figure 9: **Annotated assembly and between-individual polymorphisms.** (A) Shows a representation of the mitochondrial genomes assembly. Polymorphic sites are indicated in green with the alleles ordered as in (B). (B) The alignment of the six assemblies contains 18 polymorphic sites. Mismatches are highlighted.

The alignment of all six mitochondrial genomes assembled contains 18 variable sites, five of which show population-specific polymorphisms (Fig. 9B). A neighbour-joining tree shows that each population's individuals have sequences most similar to one other (see Fig. 27 in Appendix B)

3.3.3 NextSeq: Polymorphism within assemblies

Within each individual assembly created from short reads, SNPs were called with the minimum minor allele frequency set to 1 % to avoid erroneous calls due to sequencing errors. All assemblies contain numerous polymorphic sites with low to medium minor allele frequencies which are attributable to variants present in Numts (see ribbons of jitter in Fig. 7). The distributions of these allele frequencies are skewed towards 0 with maxima varying between samples. The extremes are 0.07 and 0.2 in samples N5 and N6 (corresponding to the narrowest and widest band in Fig. 7). Between samples, there is an inverse correlation between the maximal minor allele frequency and the fraction of Illumina NextSeq reads mapped to the mitochondrial reference. The more reads of one sample match the mitochondrial assembly, the lower are the SNP frequencies observed (Fig. 8). This correlation allows an estimation of the genomic proportion of Numts in *P. pedestris* (see Discussion).

In total, there are four SNPs with frequencies that fall outside the continuous distributions observed (see arrows in Fig. 7). Interestingly, individuals N1 and N3 from Le Blayeul share one such polymorphism at base pair 7567. These high-frequency variants are presumably signatures of heteroplasmy (Mao et al., 2014).

3.3.4 NextSeq: Distances between paired reads mapped – Signature of Numts

Paired reads mapped to the mitochondrial assembly could have originated from mitochondrial genomes or from Numts. The former are under strong selective constraint and should show mapped distances similar to the sequencing libraries' insert sizes, whereas Numt sequences may have been subject to insertions, deletions, or rearrangements potentially causing longer insert sizes or

discordant read orientations. Per-sample analysis of read distance distributions showed there is generally a global maximum around 400 bp representing the sequencing insert size (data not shown). There is a second (much shallower) peak above 15 000 bp resulting from mapping reads generated from circular molecules to a linearised reference. Fig. 10 shows each read pair mapped to a reference as a line connecting the reads' mapping positions. While most insert sizes are small generating blue circular outlines, there are very few read pairs with intermediate insert sizes ($>$ 1500 bp and $<$ 14 508 bp). These cause lines crossing the circles in Fig. 10. Such patterns may be caused by assembly or mapping artefacts, by rearrangements in Numt sequences, or by a combination of these. It might be argued that the former two are unlikely because the structure of animal mitochondrial genomes is highly conserved and the reads mapped were of high quality. A way to test whether these unusual insert size are caused by Numt sequences would be to sequences purified mitochondrial DNA. Interestingly, some of the lines shown cluster and some cluster patterns are shared across multiple samples. For instance, all individuals have read pairs connecting the regions around 3 and 6 kbp, all individuals of the Col de Mariaud population have connections between 3 and 10.5 kbp. The overall patterns are not population-specific (a linear discriminant analysis failed to assign all individuals to the correct populations, not shown here).

3.3.5 Mapping PacBio reads

PacBio reads generated from DNA of three individuals (P1-P3) were mapped to the mitochondrial assembly. Out of 297 899 non-redundant reads generated in total, 443 showed similarity to the mitochondrial assembly with a cumulated mapping length of 396 770 bp. Of these, most reads matched the mitochondrial reference only along a part of their length, indicating they were derived from Numts. There are only 59 PacBio reads matching full-length. For each of these reads, the number of alignment errors was compared to the distribution of alignment errors expected from per-base quality scores (phred scores). This indicated 41 are likely derived from Numts (they are too diverged to be derived from live mitochondria). Together with the reads matching part-length, diverged

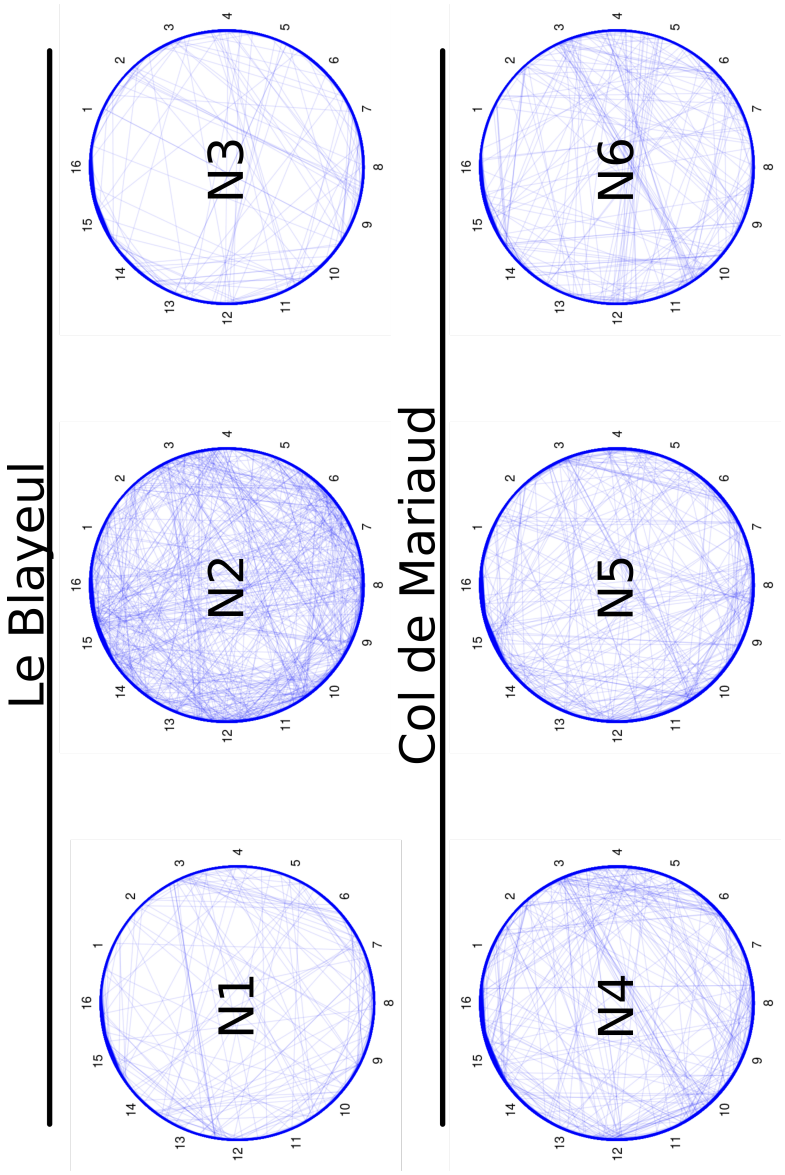

Figure 10: **Mapping position of read pairs to the mitochondrial reference.** For each of the six individuals, the mapping position of read pairs are indicated by lines. While most mates map close to one another (generating circular outlines), few pairs map far apart generating lines crossing the circles. Note common patterns between individuals. For instance, all individuals in the Col de Mariaud population show several read pairs connecting the areas at 3 and 10.5 kbp. All samples of both populations have read pairs connecting 3 and 6 kbp.

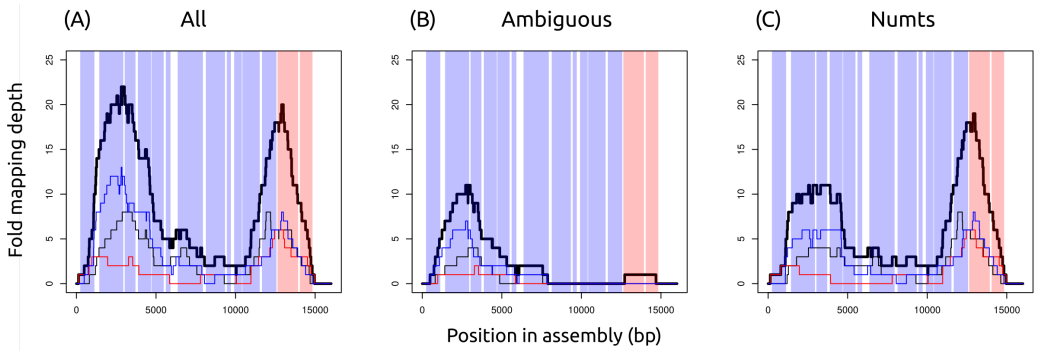

Figure 11: **Mapping depths of PacBio reads matching full-length.** The bold line indicates the sum over all four samples, narrow lines represent individual samples. Shaded areas indicate the positions of protein-coding genes with the exception of the two right-most ones which are the mitochondrial rRNA genes. (A) Shows all PacBio reads matching full-length. (B) Shows only reads whitch cannot be excluded to be derived from live mitochondria. (C) Shows reads which are derived from Numts. They are more diverged from the mitochondrial assembly than expected by the reads' error profiles (95 % confidence interval, one-tailed, Bonferroni-corrected for 59 samples).

full-length matches amount for 0.04 % of the (non-redundant) PacBio data generated. The remaining 18 full-length matches cannot be ruled out to be derived from mitochondrial genomes, but they may well be derived from Numts inserted recently. Covering 36 925 bp, these ambiguous reads represent only 9.3 % of the total length mapped to the mitochondrial assembly.

Interestingly, the mapping depth of full-length matches shows a bimodal distribution. While the 18 ambiguous matches contribute mostly to the first peak, Numt-derived reads map to the areas under both peaks (see Fig. 11).

3.4 Discussion

Animal mitochondria contain circular genomes of generally approximately 16 kbp length (Boore, 1999). Their sequences are important to many population genetic, phylogenetic, and ecological studies. Nuclear insertions of such sequences (Numts) are fascinating study objects which can be (phylo-) genetically informative (Lopez et al., 1994; Dayama et al., 2014). Conveniently,

both sequence types are often generated as side-products of other sequencing experiments similar to by-catch on fishing trawlers. But the two can be easily confused, potentially causing problems like ghost alleles and others (Zhang & Hewitt, 1996; Bensasson et al., 2001). It was found here that sequencing methods differentially enrich for one or the other kind of sequence. This comparison was done with genome skimming data (less than 10 % genomic coverage) from the grasshopper, *P. pedestris*, using Illumina's NextSeq and PacBio's RSII platforms with six and three biological replicates, respectively. Both platforms differ in many aspects such as quality of the raw data, read length, and, maybe most importantly, library preparation.

3.4.1 By-catch differs between sequencing methods

There is a clear bias towards low GC-content in the NextSeq data (see Fig. 6). This cannot be due to amplification bias (Aird et al., 2011), because the libraries sequenced here were generated PCR-free. The bias observed may be due to some inherent properties of one or both of the sequencing methods.

One of the most striking results is that mitochondrial-like RE clusters show the highest enrichment ratios for NextSeq data (16-fold at least). Mitochondrial genomes are known for very low GC-contents and these observations may be due in parts to the general bias observed towards over-representation in NextSeq data of low-GC sequences. However mitochondrial genomes are distinct from nuclear DNA, providing additional potential for bias. An important difference between the two is that animal mitochondrial genomes are only about 16 kbp long which may affect how library preparation acts on them. While NextSeq libraries are usually fragmented to sizes of 350–550 bp and subsequently size-selected to peak around these values, library preparation of long-read sequencing involves more careful shearing and a subsequent size selection for long fragments (around 3–4 kbp in this case). This may cause mitochondrial genomes to be lost from PacBio libraries while Numts, part of the nuclear DNA, may be represented in their natural proportion.

Consequently, looking at all mitochondrial-like sequences, NextSeq data show a much lower abundance of Numt sequnces (7 and 20 % per sample, Fig. 7)

than data generated with PacBio technology. The range observed is likely due to mitochondrial densities differing between samples. Using PacBio data, the proportion of Numts in all mitochondrial-like data was identified by sequence divergence and whether reads map only partially. This shows at least 90 % of the mitochondrial-like sequences obtained with PacBio technology are derived from Numts, a lower-bound estimate because ambiguous sequences may be derived from Numts with low sequence divergence.

3.4.2 The genomic proportion of Numts in *Podisma pedestris*

Building on the results presented above there are two ways of estimating the genomic proportion of Numts in the *P. pedestris* genome which are possible even in the absence of a genome assembly. It is shown in Fig. 8 that for NextSeq data there is a good correlation ($R^2 = 0.98$) between the proportion of mitochondrial-like reads and the inverse of the maximum lower allele frequencies. In other words, the more mitochondrial-like reads in a NextSeq data set, the fewer rare alleles are found (see Fig. 7, the higher the mapping coverage, the lower the SNP frequencies). Assuming that rare alleles are caused by Numts and that Numts have similar genomic proportions between samples, the genomic proportion of Numts can be calculated from the regression's intercept a and slope b. Assuming that the maximum SNP allele frequency is a function of the proportions p in the sequencing data of Numt-DNA and true mitochondrial DNA

$$freq_{max} = \frac{p_{numt}}{p_{mito} + p_{numt}}.$$

The proportion of mitochondrial-like DNA is then

$$p_{mito-like} = \frac{p_{mito} + p_{numt}}{p_{genomic} + p_{mito}}$$

Which is proportional to the inverse regression slope

$$p_{mito-like} = a + \frac{b}{freq_{max}}.$$

Setting both equations equal

$$\frac{p_{numt} + p_{mito}}{p_{genomic} + p_{mito}} = a + b\frac{p_{mito} + p_{numt}}{p_{numt}}$$

and setting $p_{mito} \to 0$ gives

$$\frac{p_{numt}}{p_{genomic}} = a + b$$

The estimate is $-0.003\,\%$. An exhaustive bootstrap results in a skewed distribution with a 95 % interval ranging from $-0.165\,\%$ to $0.101\,\%$ with the mass of estimates greater than 0 (insert in Fig. 8). This provides an upper estimate of 0.1 %.

The PacBio estimate is more straightforward. The estimate is equal to the data from Numt reads D_{numt} divided by all data minus the ambiguous data which might be mitochondrial. This gives $\frac{D_{numt}}{D_{all} - D_{ambig}} = 0.037\%$. Because of the exclusion of ambiguous reads, this is a lower-bound estimate.

3.4.3 NUMTs and Genome size

P. pedestris has the largest genome of any insect listed in the Animal Genome Size Database ((Gregory, 2016), accessed 4.04.2017). Its C-value of 16.93 corresponds to 16.5 Gbp (Doležel et al., 2003). Consequently, a genomic proportion of just under 0.1 % is equivalent to about a thousand mitochondrial genomes inserted full-length (this is comparable to a Drosphila or small Arabidopsis chromosome). A genomic proportion of this order is not uncommon. Hazkani-Covo et al. (2010) present estimates of Numt contents for a diverse list of 85 species, none of which have values greater than 1 %. RepeatExplorer analyses suggest that repeats account for approximately 70 % of the *P. pedestris* genome. It may seem surprising that mitochondria-derived DNA contributes so little to this fraction. But much of the *P. pedestris* genome's repetitive fraction is amounted for by transposable elements able to excise and re-insert themselves, providing a mechanism for copy number increase which Numts cannot exploit.

3.5 Conclusion

It is not always straightforward to determine whether a mitochondrial-like sequence is derived directly from a mitochondrial genome or from Numts. Analysing data generated with two standard sequencing approaches, it is shown here that methods complementarily enrich for one or the other DNA species. While Illumina NextSeq short reads contained four to thirteen times more mitochondrial DNA than Numts, long PacBio reads showed at least nine times more Numt sequences than mitochondrial ones. Awareness of this phenomenon will be useful when planning sequencing experiments or when analysing publicly available data.

Acknowledgements

This research utilised Queen Mary's MidPlus computational facilities, supported by QMUL Research-IT and funded by EPSRC grant EP/K000128/1. Further computational resources were provided by the ELIXIR-CZ project (LM2015047), part of the international ELIXIR infrastructure. HB is funded by a studentship of QMUL's School of Biological and Chemical Sciences.

4 How old is the split between the *Podisma pedestris* sex chromosome populations?

The grasshopper *Podisma pedestris* occurs in the Southern French Alps predominantly at elevations above 1500 m. It has two parapatric populations, which differ in their sex chromosome system. Where both populations meet, they form a narrow hybrid zone where there is evidence for strong selection against inter-population hybrids, suggesting that the zone might present a barrier to gene flow. During the Quaternary glacial cycles, *P. pedestris* would have recurrently left and re-colonised its range and the populations would have repeatedly come into contact. To explore the populations' demography, coalescent-based models were fitted to transcriptome data using the block-wise site frequency spectrum of two diploid individuals. The results suggest a split several glacial cycles ago and a present-day population size far smaller than during times of glaciation.

4.1 Introduction

The genetic make-up of Europe and North America's biodiversity has been greatly shaped by the Quaternary climatic cycles. The alternating glacial maxima and interglacials made species oscillate between northerly or high-altitude habitats and more southerly or lower-altitude refugia (Hewitt, 2000). Genetic drift during times of allopatry and drift-related evolutionary processes during phases of relocation such as the effect of long distance dispersal (Nichols & Hewitt, 1994), embolism effect (Bialozyt et al., 2006), bottlenecking, and allele surfing (Klopfstein et al., 2006) may have occurred providing opportunity for populations to diverge. If diverged populations evolved partial reproductive isolation (they may have changed their mating behaviour, adapted to different ecological optima, or evolved intrinsic reproductive isolation to some extent), they will on secondary contact form hybrid zones, where gene flow may happen through hybrids. Hybrids will then be less fit or may be confined to a specific ecotone posing a barrier to gene flow. Mountain ranges may provide both,

ecotones and physical barriers to gene flow, and so it is not surprising that many hybrid zones are associated with mountains (Swenson & Howard, 2005; Hewitt, 2011).

One species with such a hybrid zone is the flightless Alpine grasshopper, *Podisma pedestris*, which is limited to mountain ranges in Central Europe. While the species has a X0 sex chromosome system throughout most of its range, a neo-XY population exists in the Southern French Alps where the grasshopper occurs above 1500 m mainly. Where both populations meet, they form a hybrid zone with a narrow cline for the transition between the sex chromosome systems (Hewitt, 1975). In this zone, strong selection is observed against hybrid offspring (Barton, 1980; Barton & Hewitt, 1981b; Nichols & Hewitt, 1988), which seems to suggest that the hybrid zone is a strong barrier to gene flow. The fact that the hybrid zone follows a major watershed (see Chapter 2) lends support to the idea that it has arisen in secondary contact. This does not necessarily mean the hybrid zone is the result of the first re-encounter of the populations. Following the Quaternary climatic cycles, the populations may have split and re-joined several times before (Nichols & Hewitt, 1986), a setting of para-allopatry. While there are verbal models, no molecular data has so far been used to explore the species' demography. This is the purpose of this chapter.

With a C-value of approximately 17 pg ((Westerman et al., 1987), about 100 times the value of *Drosophila melanogaster*), *P. pedestris* is the insect with the largest genome size reported (`http://www.genomesize.com`). Making sequencing expensive and complicating genome assembly, this poses a serious challenge to genomic analyses. To circumvent these issues, the demographic analyses carried out here are based on transcriptome data. Following the approach of Lohse et al. (2011), polymorphism data from equally-sized blocks of sequence (the block-wise site frequency spectrum) are used to fit demographic models.

Felsenstein's (1988) equation can be used to compute the likelihood of genetic data considering all possible genealogies (each weighted by their probability). Because there are only four lineages sampled here (two diploid transcriptomes), the probabilities of all possible genealogies can be computed per locus and can then be multiplied over all loci (as done, for instance, by Takahata et al. (1995)). This yields the likelihood of a demographic model. These models

Table 3: **Sample information of the two individuals sequenced.**

Karyotype	**neo-XY**	**X0**
Location	Le Blayeul	Col Bas
Read pairs generated	40 750 427	42 596 157
Latitude	44.265 799 N	44.384 855 N
Longitude	6.305 988 E	6.400 470 E
Trinity isoforms	190 329	173 166
Trinity genes	110 951	105 356
Per-block π	0.36	0.52
Per-block d_{xy} (= π_{12})	2.51	

are represented by generating functions of genealogies. The probability of a particular configuration of mutations can be obtained by differentiating the generating function. The inferred model and parameter estimates are discussed regarding the European glaciations and the history of the neo-XY chromosomes.

4.2 Materials and methods

4.2.1 Sample collection and sequencing

Hind legs were removed from two *P. pedestris* individuals (see Tab. 3). They were snap-frozen inside plastic tubes, in a mixture of white spirit and solid carbon dioxide, and were then stored at $-79\,^{\circ}\mathrm{C}$ until shipping for which they were transferred into RNAlater with the cuticle cut open to enable more rapid penetration of the tissue. RNA extraction, library preparation, and direction-specific sequencing were carried out by HudsonAlpha (Huntsville, Al, USA). RNA-libraries were ploy-A selected, but not normalised.

4.2.2 Transcriptome assembly

Raw transcriptome reads were checked using FastQC (version 0.11.4). This identified numerous reads with adapter sequences of low quality at the 3'-end. Skewer (version 0.2.2 (Jiang et al., 2014)) was then used to trim each read's 3'-end until the read's average phred quality was ≥ 30. Trinity (trinityrnaseq_r20140717

(Grabherr et al., 2011)) was used to assemble both transcriptomes separately using the command line:

```
Trinity --seqType fq --JM 100G --left <left_reads.fastq>
    --right <right_reads.fastq> --SS_lib_type FR --CPU 10
    --full_cleanup_ET --min_kmer_cov 2  --group_pairs_distance 400
    --min_per_id_same_path 98 > <log_file>.log
```

Transcript Quantification was then carried out using the script `align_and_estimate_abundance.pl`, which is supplied with the Trinity package. Only contigs with read support (FPKM>0) were kept for further analysis.

4.2.3 Paralogue filtering and variant calling

The output of Trinity consists of sequences clustered in a three-level hierarchy. Each sequence is called an isoform, several isoforms belong to a 'gene', and several 'genes' can form a component. Isoforms may be different products of the same actual gene (splice variants), but they may also be transcripts derived from several separate paralogous genes (or they may contain paralogous exons). To exclude potential paralogues from the analysis, paralogue filtering was carried out following the method explained in the supplementary data of Nürnberger et al. (2016), outlined here:

In both assemblies, coding regions were predicted and translated using TransDecoder (version 3.0.2, https://github.com/TransDecoder/TransDecoder). Paralogue detection was then carried out on the level of Trinity 'genes'. For each gene, the isoform containing the longest ORF was labelled as reference (if there were multiple longest ORFs, one was chosen randomly). Subsequently all isoforms per Trinity gene were aligned to their reference. Alignments were split into high and low-quality blocks. Isoforms were deemed potential paralogues of their reference if sequence identity in high-quality blocks was lower than 98 %, or if sequence identity in low-quality blocks was greater than 50 %. The rationale behind this is that there will be alignments of true isoforms with different exons expressed. Such alignments will necessarily contain stretches of low-quality which is not a reason to discard them as paralogous. It is also possible however,

that individual exons underwent duplication creating paralogous copies. If their sequences have not diverged, they will not be detected and will not have an effect, but diverged paralogous exons might interfere with down-stream analyses. Alignments of such paralogous exon sequences might show relatively high-identities but are expected to be poorer than orthologous sequences (from homologous chromosomes). In the absence of a genome assembly it is difficult to assess the duplication status of genes. The quality thresholds chosen here are conservative (they are likely to falsely exclude true some true orthologues, but most sequences deemed orthologous are likely to be true). The thresholds are not species-specific.

OrthoFinder (latest version retrieved from GitHub on 23.11.2016 (Emms & Kelly, 2015)) was then run on a joint set of peptide sequences of each reference isoform, their identified paralogues, and peptide sequences of *Locusta migratoria* (generated using TransDecoder from assembled transcriptomes retrieved from GenBank: GBDZ00000000 (Liu et al., 2014) and GDIO00000000 (Tu et al., 2015)). Of the clusters obtained, only those were kept that contained at least one sequence from each *P. pedestris* sample sequenced but none of the paralogues identified previously. Orthologue clusters with sequence similarity to non-Metazoan targets (as identified by local blastn against a copy of the NCBI nucleotide database) were removed to avoid contaminations. Suitable peptide alignments were then back-translated using PAL2NAL (version 14.0 (Suyama et al., 2006)).

Alignments generated with bowtie2 (version 2.3.0) were then used for variant detection with GATK (version 3.7 (McKenna et al., 2010) following the package's best practices recommended for detecting 'germline variation' (DePristo et al., 2011). In brief, variant sites were called preliminarily, their qualities were adjusted, and variant calling was repeated.

4.2.4 Block-wise coalescent analysis

As an extension of the site frequency spectrum, alignments can be split into blocks of equal length in which site types are counted. The numbers of blocks with the same site configurations are then tabulated to give the block-wise site

frequency spectrum (bSFS) which contains more information than the SFS (it contains linkage information). These blocks are treated as loci within which no recombination occurs and which are unlinked to one another. There is a trade-off when choosing the block size. Long blocks contain more information which will provide more power. But the longer the block size the fewer blocks will be found and the more likely it becomes that recombination is observed within blocks. Because there are comparatively few blocks compared to the *P. pedestris* genome size, blocks can be assumed to be completely unlinked. For the analyses carried out here a block length of 120 was chosen, a compromise between per-block variability and total number of blocks.

The VCF files generated with GATK as described above were then parsed using custom-made perl scripts provided by Beate Nürnberger. These had been used for a similar analysis in the *Bombina* hybrid zone (Nürnberger et al., 2016) in order to generate so-called alternate haplotype sequences (these are not truly phased). The alternate haplotype sequences were then aligned across samples and variants in blocks of 120 consecutive four-fold degenerate sites were recorded. Selection on four-fold degenerate sites was assumed to be negligible, a precondition for the coalescent-based analysis carried out here.

In alignments of sequences from two diploid individuals, A and B, there are four possible site types: heterozygous in A, heterozygous in B, shared heterozygous sites, and individuals fixed for different alleles. These types were recorded per block generating the block-wise site frequency spectrum (bSFS).

A set of utility functions implemented in Mathematica (published as supplementary data to (Lohse et al., 2016)) was used to fit two sets of demographic models to bSFS data. Models were represented by generating functions of branch lengths. Differentiating such a generating function and multiplying over all loci, it is possible to compute the likelihood of demographic models under the coalescent framework (Lohse et al., 2011). Maximum likelihood estimates (MLEs) were obtained using Mathematica's function `NMaximise`, which makes use of the Nelder-Mead method (a heuristic which is susceptible to starting values and bounds chosen).

Two classes of models were set up, divergence (DIV) and isolation with migration (IM). Each class assumes two daughter populations corresponding

to the sex-chromosomal populations which, going backwards in time, merge into one ancestral population T units of time ago (measured in units of 2Ne). While the populations are isolated in DIV, IM assumes constant unidirectional gene flow. In this framework, the effect of different population sizes can be set via a scaling factor of the coalescence rate, λ. λ is inversely proportional to the effective population size and can be set for each population in the model. λ_{anc} of the ancestral population was set to 1, providing a reference value for the daughter populations. In the models DIV_{add} and IM_{add}, λ-values were constrained so that

$$\frac{1}{\lambda_{X0}} + \frac{1}{\lambda_{XY}} = 1.$$

This means the two daughter population sizes add up to give the ancestral one. For technical reasons a grid search had to be performed across a set of fixed vales for λ_{X0}. These were 0.70, 0.67, 0.625, 0.57, 0.5, 0.43, 0.375, 0.33 and 0.3. For the DIV model it was possible to fit independent λ-parameters for each daughter population. The model variant is called DIV_{free}. The generating function for the analogous model IM_{free} allowing gene flow and independent daughter population sizes proved too complex to obtain an MLE (execution of the function `NMaximise[]` resulted in errors). A different set of generating functions was set up modelling two and three episodes of discrete admixture. For these generating functions, too, ML estimation failed due to technical limitations. All models are shown in Fig. 12.

In order to obtain a measure of the uncertainty of the MLE, a bootstrap (100 replicates) was performed based on the parameters fitted to the DIV_{free} model. One hundred sets of 740 blocks were sampled from the expected bSFS given the parameters estimated in DIV_{free} and the model was fitted to these. The parameter distributions obtained are shown in Figures 15 and 16. The 0.025 and 0.975 quantiles are given in Tab. 5.

4.2.5 Biological relevance of the models investigated

All models fitted here assume that the range of an ancestral X0-population is now occupied by an X0 and a derived XY-population. The models with additive effective population sizes (subscript "ADD") somewhat naively assume that

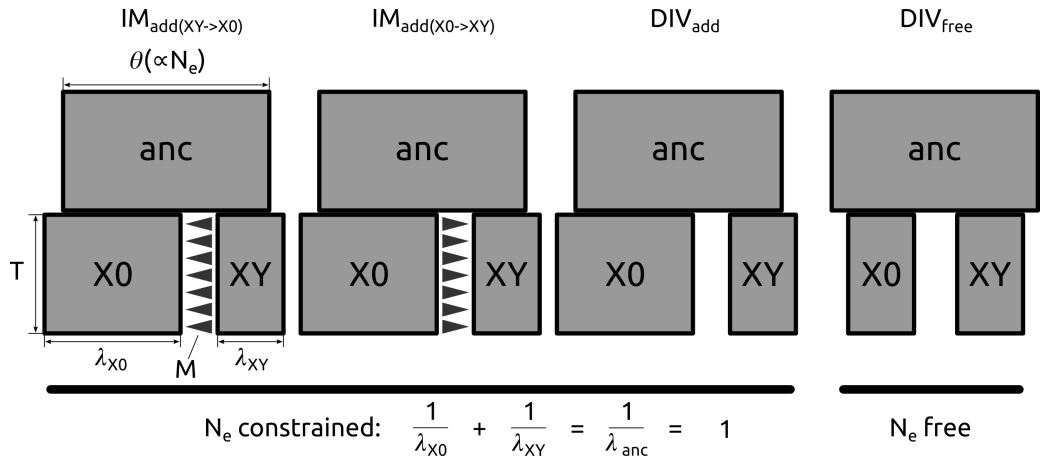

Figure 12: **Demographic models fitted.** An ancestral population splits instantaneously into two daughter populations. Models with subscript 'add' have the daughter populations' sizes constrained so that they add up to the ancestral one.

the effective population sizes observed at present add up to give the ancestral one. Because effective population size does not necessarily correspond with geographical range (or census population size) size, $\mathrm{DIV}_{\mathrm{free}}$ was fitted, where all effective population size parameters were fitted independently (at the cost of estimating one more parameter). Thus, only the additive models allow to estimate a gene flow parameter.

A different way of implementing gene flow is to assume discrete pulses of gene flow at certain times (corresponding to presumed population expansions in the interglacials).

4.3 Results

4.3.1 The data

Two *de novo*-transcriptome assemblies were generated: one of an XY-individual with 190 329 isoforms in 110 951 components and one of an X0-individual with 173 166 isoforms in 105 356 components. Clustering the isoforms with Orthofinder (Emms & Kelly, 2015) resulted in 15 810 orthogroups (groups of potentially orthologues sequences). After removal of sequences deemed

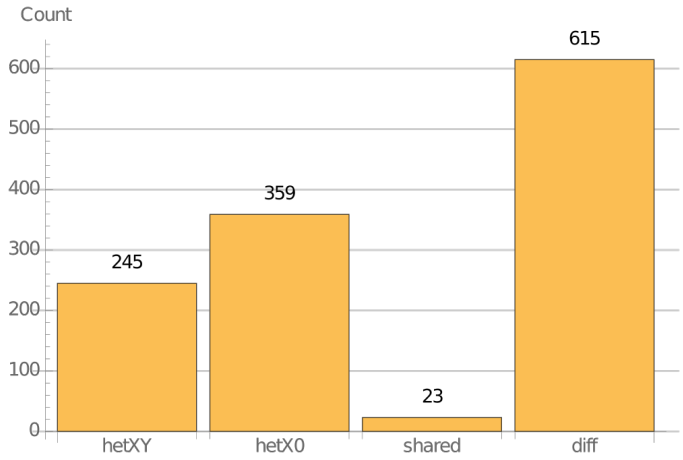

Figure 13: **Site-frequency spectrum obtained from 740 blocks of 120 consecutive four-fold degenerate sites between two diploid transcriptomes.** The average number of variant sites per block is $S_k = 1.68$.

paralogous and selection sequences of sufficient length following the method of (Nürnberger et al., 2016), 740 sequence blocks of 120 consecutive four-fold degenerate sites were identified. The site frequency spectrum observed is shown in Fig. 13. Other summaries can be found in Tab. 3.

4.3.2 Models fitted

Assuming that the ancestral range of *P. pedestris* is now shared between two parapatric populations differing in their sex-chromosome system, models were fitted initially with the daughter population sizes constrained to add up to the ancestral population size. Two models, DIV_{add} without unidirectional gene flow and IM_{add} with unidirectional gene flow were fitted. The parameters fitted to the DIV_{add} model are shown in Tab. 4. The highest log likelihood was obtained for a population size ratio between X0 and neo-XY of 58/42 with the scaled mutation rate θ (per block) of 0.663 and a split time of $T = 0.693$ (in units of $2N_e$ generations).

The best-fitting parameter sets of the IM_{add} models are shown in Tab. 4 as well. While the best estimates for theta are similar to the DIV_{add} model, the presence of gene flow causes older split times to be fitted. The optimal population size ratio depends on the direction of gene flow assumed. The receiving populations

Table 4: **Parameter estimates.** In the left-hand three models, the relative population sizes of XY and X0 add up to one. The difference in log-likelihood is given to the best-fitting model DIV_{free}. θ is the scaled mutation rate ($\theta = \mu 4N_e$). The split time T is given in units of $2N_e$. M is the scaled migration rate ($M = m4N_e$).

	DIV_{add}	$\text{IM}_{\text{add(X0}\rightarrow\text{XY)}}$	$\text{IM}_{\text{add(XY}\rightarrow\text{X0)}}$	DIV_{free}
rel N_e of X0	0.58	0.625	0.5	0.58
ΔlnL	-8.8	-7.7	-7.8	0
θ	0.663	0.642	0.61	0.91
T	0.693	0.827	1.05	0.325
M	n/a	0.287	0.503	n/a

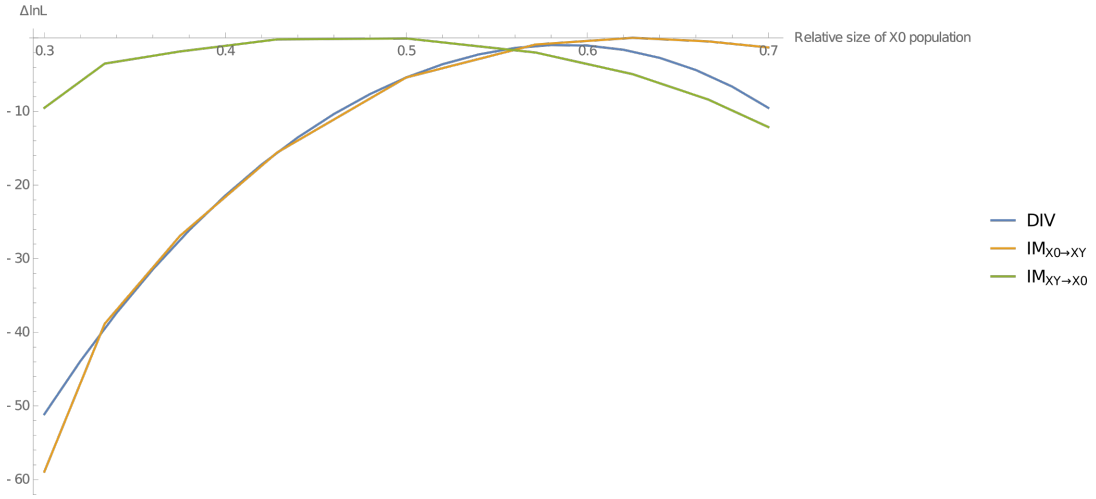

Figure 14: **Marginal likelihoods for several population size combinations of the additive versions of the DIV and IM models** plotted against the relative size of the X0 population. The colours represent: blue – no gene flow, mustard – gene flow into XY, green – gene flow into X0.

were generally fitted a smaller (relative) effective population size than in the DIV_{add} model.

All models have very similar log-likelihoods. The likelihood ratio between the worst and the best is 1.1. A likelihood ratio test shows that the difference is not significant $p(\chi^2_{2.2,df=1}) = 0.138$.

Table 5: **Parameters fitted to the $\mathrm{DIV}_{\mathrm{free}}$ model** and 95 % confidence intervals based on 100 parametric bootstrap replicates.

	2.5-percentile	$\mathrm{DIV}_{\mathrm{free}}$	97.5-percentile
lnL	−2027	−1978	−1847
θ	0.7878	0.9085	1.151
T	0.1499	0.3261	0.473
λ_{X0}	2.465	3.435	6.833
λ_{XY}	3.148	4.672	9.704
age of the split	220 000	411 000	603 000

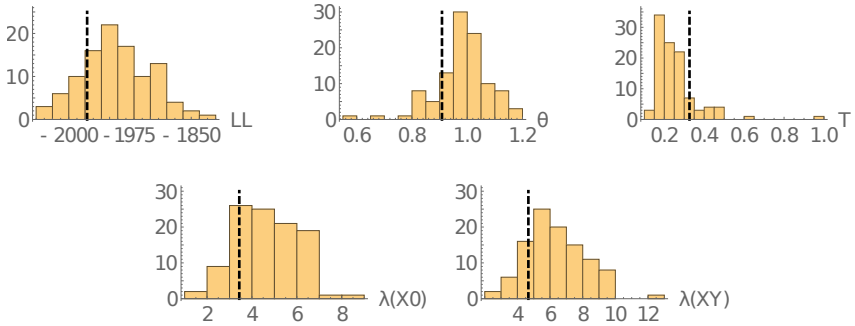

Figure 15: **Distribution of model parameters estimated from 100 bootstrap replicates for the $\mathrm{DIV}_{\mathrm{free}}$ model.** The dashed lines indicate the original model estimates.

To test whether fits improve when both daughter population sizes are fitted in an unconstrained fashion, $\mathrm{DIV}_{\mathrm{free}}$ was set up where the current-day population sizes could vary freely. The fitted population sizes are approximately 21 % for the neo-XY and 29 % for the X0 population, however these have wide confidence intervals (see Tab. 5). An analogous model was set up with gene flow, $\mathrm{IM}_{\mathrm{free}}$, but the generating function was too complex to obtain an MLE.

4.4 Discussion

4.4.1 Modelling a complex population history

The demographic models investigated here assume two present-day populations (one corresponding to each of the *P. pedestris* sex-chromosome configurations),

which were derived from an ancestral population in an instant split. While it is assumed that the *P. pedestris* populations will have met during comparatively short interglacials (Nichols & Hewitt, 1986) possibly giving opportunity for pulses of admixture, the models implemented here represent two extreme situations: divergence without gene flow, and divergence with subsequent directional gene flow between populations. It was attempted to model potentially more realistic situations allowing for discrete pulses of admixture after the split of the ancestral populations, but maximum likelihood estimation failed in these more complex models. Considering the species' comparatively low dispersal ability (the standard deviation of per-generation dispersal was estimated as 20 m (Barton & Hewitt, 1982)) it may not be expected to detect signatures of gene flow. However, the DIV_{add} model without and the IM_{add} model with gene flow fit the data similarly well (a likelihood ratio did not find a significant difference). It would be desirable to carry out a power analysis to assess whether a clearer answer would be possible using more data. For example, the expected block-wise site frequency spectrum under the parameters of the best-fitting IM_{add} model could be used to generate parametric bootstrap samples in order to fit the DIV_{add} model to these. Such an analysis was attempted, but unfortunately model fitting failed for most bootstrap replicates for technical reasons. In particular, maximising likelihood with `NMaximise[]` is very susceptible to the parameter bounds chosen. While these bounds can be adapted for individual runs, this is not reasonable for bootstrap replicates.

Nevertheless, the gene flow parameter estimate of the IM_{add} model is $M = 0.287$ which corresponds to one migrant every 4 generations. This is about one fifth of the value inferred for the sympatric butterfly sister species *Heliconius melpomene rosina* and *H. cydno* (Lohse et al., 2016), and it is one order of magnitude higher than the value inferred for *Bombina bombina* and *B. variegata variegata*, two ecologically diverged, para-allopatric species of toad (Nürnberger et al., 2016). In light of the results that will be obtained in Chapter 5, the presence of gene flow seems not unlikely.

Models with additive population sizes (denoted by the subscript 'add') had been set up to reflect the idea that the ancestrally continuous range of *P. pedestris* is now split between two chromosomal populations. This class of models tends

to fit a lower size to the neo-XY population which indeed has a smaller range. Albeit, the populations size ratio fitted (58/42 in the DIV_{add} model) is much less extreme than the comparatively small range of the neo-XY population would suggest. If gene flow is allowed into the X0 population (IM_{add}), the population sizes are even inferred to be equal. However, considering the limited dispersal capabilities of *P. pedesris*, it is not surprising that range size is a bad predictor of the effective population size and that the species shows considerable structure within the chromosomal populations (as will be shown in Chapter 5). This suggests that constraining the daughter population sizes is not necessarily very realistic. The model DIV_{free} accounts for that, allowing both daughter population sizes to be fitted independently. The DIV_{free} model showed a significantly better fit than the additive models (8 log likelihood units $p(\chi^2_{16,df=1}) = 6.3 \times 10^{-5}$. The population size estimated for the neo-XY and X0 populations are 21 % and 29 % of the ancestral population size. The ratio of these values is very similar to the ratio of per-individual values of π (see Tab. 3). In a population of constant size, with constant mutation rate, and under neutrality, π is proportional to the effective population size (see below). The fact that π and population size estimates agree can be seen as a sanity check for the model.

Given a per-generation mutation rate μ, the effective population size can be calculated from theta via the relationship $\theta = 4N_e\mu$. The mutation rate of *P. pedestris* is unknown, but recent estimates for two other insects, *Heliconius melpomene* (Keightley et al., 2015) and *Drosophila melanogaster* (Keightley et al., 2014) are very similar and close to $\mu = 3 \times 10^{-9}$ per generation and base pair. While the models with additive population sizes estimate similar values ranging around 450 000 years, DIV_{free} gives an estimate of 630 000 years.

During periods of glaciation when *P. pedestris* would have lived at lower elevations, the species may have had a range much larger and perhaps less topographically-structured than today (Nichols & Hewitt, 1986). This is well reflected in the model DIV_{free} which shows the best fit to the data and where both daughter population sizes are small compared to the ancestral one.

4.4.2 The age of the neo-XY population

Two populations of *P. pedestris* meet in a hybrid zone, one has the ancestral X0 sex chromosome system, the other carries neo-XY chromosomes. It seems plausible that they have recurrently come in contact and split again following the Quaternary glacial cycles (Nichols & Hewitt, 1986), a para-allopatric system. As pointed out in Chapter 2, it is likely that the ranges of the chromosomal populations have not changed over the last 10 000 years relative to the watershed which separates them today. It seems obvious to ask how old is the split between the two, how old is the neo-XY population? But what actually defines this population? John & Hewitt (1970) who first described the neo-sex chromosomes suggested the neo-X goes back to a single X-to-autosome fusion, which must have originated in a small, subsequently expanding population. In such a scenario, all individuals carrying the neo-sex chromosomes should share a similar genomic background and the chromosomal fusion would be as old or nearly as old as the split between both populations. An alternative idea is the inexorable spread phenomenon first suggested by Veltsos et al. (2008): Sexually antagonistic selection is likely to be acting on the neo-Y chromosome which has an area of reduced recombination (John & Hewitt, 1970; Hewitt & John, 1972). Due to their deleterious effect in hybrid females, the presence of neo-Y chromosomes would give an indirect advantage to the large neo-X chromosome (females homozygous for the neo-X cannot carry a neo-Y). This would cause the spread of the neo-sex chromosomes into the ancestral genomic background. In this case, grasshoppers carrying the neo-X might have different genomic backgrounds only sharing the fusion centromere and the sex-chromosomal region of the neo-Y which cause the inexorable spread phenomenon (loci less tightly linked might have have been lost through recombination). If this is the case, the age of the chromosomal fusion might be very different to the split inferred from genomic data. However, there is not much evidence that inexorable spread had a major influence on the *P. pedestris* sex-chromosome distribution (see Chapter 2).

Using the effective population sizes computed above, the relative split times T can be translated into a number of generations (which equals time in years

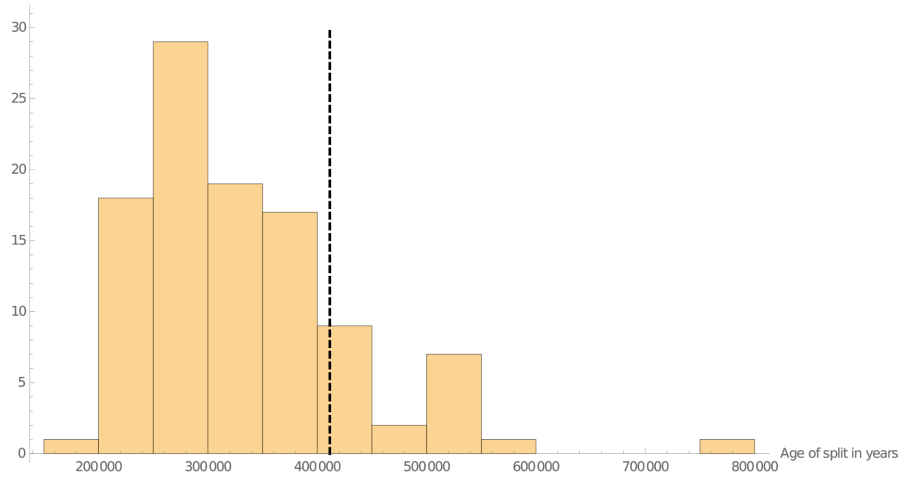

Figure 16: **Distribution of split times determined from 100 bootstrap replicates of the $\mathrm{DIV}_{\mathrm{free}}$ model.** The dashed line indicates the point estimate.

for the univoltine species *P. pedestris*). These times vary between the models by a factor of approximately 2. All fall into the Pleistocene. While the estimates of the models with additive populations sizes are older (640 000 years or more if gene flow is allowed). The estimate of the $\mathrm{DIV}_{\mathrm{free}}$ model is approximately 411 000 years. The bootstrapped split values are centred around 275 000 years (see Fig. 16) suggesting the populations split at least two glacial cycles ago. However as discussed above, is is not unlikely that homogenising gene flow has occurred and that the split is considerably older.

4.4.3 Possible sources of error

There are some likely sources of error which may have affected this study. Analysing polymorphism data from four-fold synonymous sites, it was implicitly assumed that selection can be ignored. But since these sites analysed are parts of genes, they will certainly be subject to some kind of linked selection. For instance it is common in bSFS analyses to find the class of blocks without mutations to be over-represented compared to the expectation of the best-fitting model. This was the case here, too (data not shown), which is likely the signature of background selection removing variability from linked sites. Selective

sweeps would have a similar effect. Another assumption made was that blocks are unlinked and that there are infinite sites. There were however two blocks which violated this assumption, showing sites with both fixed differences and shared alleles. Such configurations are only possible as the result of recurrent mutation (violating the infinite sites assumption) or when recombination or gene conversion occur. The blocks were removed from the analysis. Other sources of error include the possible heterogeneity of mutation rates across the genome and genetic structure in the populations sampled. For instance F_{ST} can be calculated from the values of π and d_{xy} given in Tab. 3. The resulting value of 0.82 is much higher than the one obtained in Chapter 5 from multi-individual data from Dormillouse (which is near the sampling site of Col Bas) and from Col de Mariaud. Another possible error source is allele-specific expression (Degner et al., 2009) from which variant calling in transcriptome data is prone to suffer. Allele-specific expression might cause fewer heterozygous sites to be detected, potentially influencing inference in complex ways.

4.5 Conclusion

The species studied here, *Podisma pedestris*, is a flightless grasshopper which in the Southern Alps is confined to altitudes mainly above 1500 m. Several factors are likely to have strongly affected the evolution of its genetic make-up: (1) Its limited dispersal (Barton & Hewitt, 1982), (2) its Alpine range covers a spatially highly structured habitat, and (3) during the Quaternary climatic cycles the species underwent recurrent cycles of retreat from and re-colonisation of the Alps (Nichols & Hewitt, 1986).

Of several demographic models investigated, DIV_{free} (divergence without gene flow and free daughter population sizes) fitted the data best suggesting a split time approximately 411 000 years ago (over three glacial cycles ago). Considering gene flow may be occurring, this can be taken as a lower-bound estimate. The model also fits comparatively small present-day population sizes, likely reflecting that the current-day range is much smaller than the one inhabited during glacial maxima.

5 The genetic architecture of the hybrid zone

Two parapatric populations of the grasshopper *Podisma pedestris* differ in their sex chromosome system, one carrying an X0 and the other one a neo-XY system. Where they meet they interbreed and strong selection is observed against hybrids. The X-chromosomal cline is far wider than the observed selection would suggest, which led to the idea that selection against hybrids is due to hundreds of loci of small effect acting together (Barton & Hewitt, 1981b). Here the genetic structure of the hybrid zone is investigated using some 10 000 single-nucleotide polymorphisms from hundreds of genomic loci replicated over several transects crossing the hybrid zone. Fitting clines, only two sites are found showing patterns of selection across transects. Both are located on the X chromosome. The findings are discussed regarding population structure, the species' colonisation history and a power analysis is performed.

5.1 Introduction

Identifying the mechanisms causing reproductive isolation between groups and uncovering how these mechanisms originated are the two main subjects of speciation research (Coyne & Orr, 2004). Both are extraordinarily complex tasks. (Coyne & Orr, 2004) provide a classification system for forms of reproductive isolation: they can be prezygotic or postzygotic of which the prezygotic class can be subdivided further depending on whether isolation acts premating or postmating. Post-zygotic isolation is probably the easiest one to invoke. It will work without any need for selection. When a population is subdivided and the daughter populations evolve in allopatry, their genes will inevitably diverge due to genetic drift. After a long-enough time they will have become incompatible.

Hewitt (2004) argued that a major driver of population subdivision and of divergence in allopatry were the Quaternary climate changes, which made species move recurrently between more northerly and more southerly geographical ranges and, in any one locality, between higher and lower altitude.

The particularities of European geography influence the consequences of these movements, the South is split into several peninsulas, and also contains areas of high relief. For many species this means their range becomes split as they move south, and also that the southern populations are more persistent – because they can retreat to high altitude as the climate warms Hewitt (2004).

The flightless Alpine grasshopper *P. pedestris* provides an excellent example of these processes. It has a hybrid zone in the Southern French Alps where two parapatric populations interbreed, which differ in their sex chromosome systems. Their hybrid progeny show severely reduced fitness (Barton, 1980). These populations likely met shortly after their current-day habitat became ice-free after the last glacial maximum, as proposed by (Nichols & Hewitt, 1986); an idea re-evaluated in Chapter 2. There are no known differences in ecological preferences between the two populations (Nichols & Hewitt, 1988), assortative mating is not observed, but there is some evidence of postmating prezygotic isolation (Hewitt et al., 1987). The sex-chromosomal cline, at 800 m, is much wider than selection acting against hybrids inferred from inter-population crosses would suggest (Barton, 1980), and the zone of hybrid inviability (400 m) in nature is much wider than the species' dispersal per generation (estimated to have a standard deviation of 20 m (Barton & Hewitt, 1982). This discrepancy between zone width and dispersal led to the suggestion that the inviability is due to several genomic loci of small effect, the result of two diverged genomes coming together. Using a maximum likelihood approach the number of loci involved was estimated to be 150 (Barton & Hewitt, 1981b).

More than 35 years have passed since this last publication on the genomic architecture of the isolation observed in *P. pedestris* and more recent results prompt re-evaluation of this view. Several studies have not found genetic differentiation across the hybrid zone Halliday et al. (1983, 1984) and others have presented evidence for introgression of rDNA with different sequences (Keller et al., 2008) and occurring at different loci (Veltsos et al., 2009).

Theoretical considerations also suggest problems with a model comprising multiple selected loci in a hybrid zone. Unless there is a high rate of recombination between the loci, they would generate stepped clines (clines with a rapid transition at the centre and shallow tails, (Barton, 1983)), a pattern not normally

found in the *P. pedestris* hybrid zone (Barton & Hewitt, 1981a; Nichols & Hewitt, 1986). In addition, evidence that homozygous females (both X-chromosomes fused, or both unfused) produce an excess of embryos with the same karyotype, suggests some form of postmating prezygotic isolation Hewitt et al. (1987, 1989).

In order to re-evaluate the number and location of genetic differences responsible for these effects, a genome-wide survey was designed. Because of the species' extraordinary genome size (approx 20 Gbp), complete genomic sequencing and assembly is impracticable, even with the latest technology. Instead, transcriptome sequencing was carried out to infer gene sequences, which were then used to design baits for a targeted sequence capture experiment. Such a reduced-complexity experiment is very unlikely to identify the loci on which selection is acting. Because of linkage, however, loci in the proximity of selected ones are expected to show similar patterns. The sequence capture process led to the successful targeting of hundreds of genomic regions containing some 10 000 SNPs. The samples were selected from several replicate transects crossing the hybrid zone to tease apart patterns generated by random neutral processes, from those generated by selection. The genetic transitions in each transect were characterised by fitting sigmoid clines to the allele frequency change. The width of a cline can be used to infer patterns of selection acting in the hybrid zone (Barton & Gale, 1993).

Following the idea proposed by Barton & Hewitt (1981b) of underdominance at numerous loci across the genome, it would be reasonable to expect five to several dozen loci showing clines narrower than 5 km, see power analysis below.

5.2 Materials and methods

5.2.1 Power analysis

In order to interpret the results that will be presented below, it is necessary to know what number of loci under selection the experiment carried out here would be expected to identify. This power analysis will assume that selection against hybrids of $s_{total} = 0.25$ is caused by $n = 150$ or $n = 500$ loci with equal selection coefficients acting multiplicatively. 0.25 is in the lower range of what was found in crossing experiments (Barton, 1980; Barton & Hewitt, 1981b). The

assumption of equal contribution of loci is a simplification. If selection differs between loci, those with higher selection coefficients will cause a stronger signal of genetic draft and will be more likely to be detected, while those subject to weaker selection will less likely be detected. Informed by the numbers of cross-overs observed on average, the total map length of the *P. pedestris* genome will be assumed to be $R = 9.3$ as by Barton (1983).

Selected sites have an effect of genomic loci nearby: this is linked selection. The effect of linked selection will decrease over time as recombination breaks the non-random association between loci (linkage disequilibria). The length of a linked block is approximately $2\sqrt{\frac{s}{t}}$ (Barton, 1979a). If the selection acting is equally spread across multiple loci acting multiplicatively, the per-locus selection can be obtained by solving the equation $1 - s_{total} = (1 - s)^n$. For 150 and 500 loci, this gives per-locus selection coefficients of $s_{150} = 0.0019$ and $s_{500} = 0.00058$. At equilibrium such selection against heterozygotes would cause cline widths of $w = 2\sigma\sqrt{\frac{2}{s}}$. Using $\sigma = 20\,\mathrm{m}$ (Barton & Hewitt, 1982) yields $w_{150} = 1570\,\mathrm{m}$ and $w_{500} = 2370\,\mathrm{m}$. The fraction of the genome showing the effect of linked selection is $\frac{2n}{R}\sqrt{\frac{s}{t}}$. The number of marker loci which will lie in this region can be assumed to be binomially distributed with $p = \frac{2n}{R}\sqrt{\frac{s}{t}}$. The distributions are shown in Fig. 17. The 95 % limits around the mean are 5 and 18 for 150 loci or 12 and 30 for 500 loci.

5.2.2 Transcriptome sample handling and sequencing

Samples for transcriptome sequencing are listed in Tab. 6. After dissection, tissues were placed in plastic tubes (VWR, Lutterworth, England) and snap-frozen by immersing the tube in a mix of solid carbon dioxide and white spirit. Samples were then stored on dry ice at $-79\,^{\circ}\mathrm{C}$ until shipping. Before shipping at room temperature to the sequencing facility, frozen tissues were transferred into RNAlater. RNA extraction, direction-specific library preparation, and sequencing was carried out by HudsonAlpha (Huntsville, Al, USA).

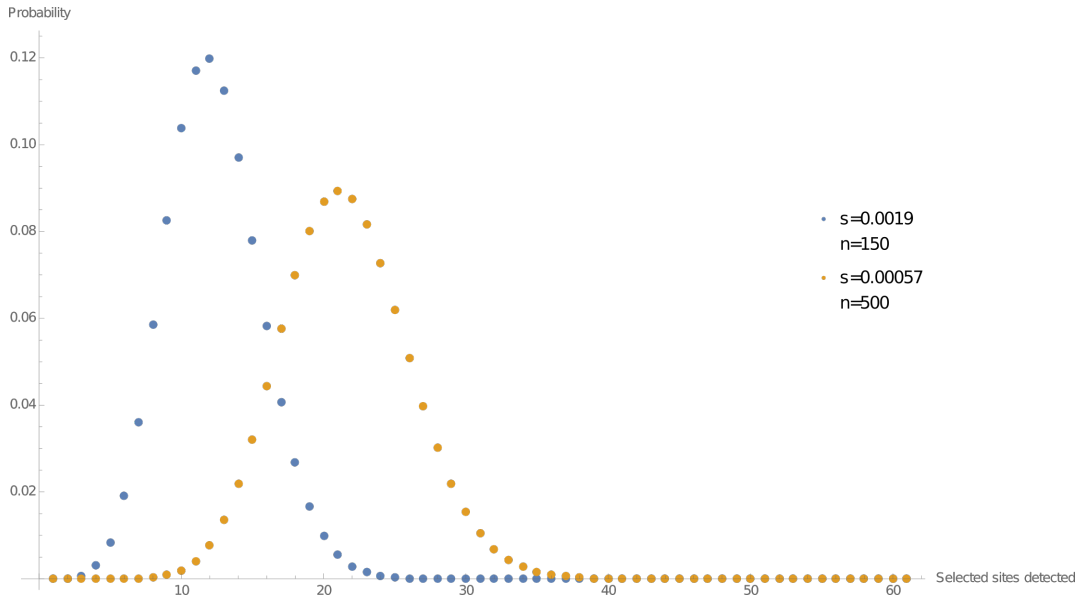

Figure 17: **Distributions of the expected number of selected loci to be detected.** For selection against hybrids of 25 % distributed multiplicatively over 150 or 500 loci.

5.2.3 Transcriptome assemblies, marker choice, and bait design

Raw transcriptome reads were checked using FastQC (version 0.11.4). This identified numerous reads with adapter sequences of low quality at the 3′-end. Skewer (version 0.2.2 (Jiang et al., 2014)) was therefore used to trim each read's 3′-end until the read's average phred quality was ≥ 30. Six individual *de novo* assemblies were generated using Trinity (trinityrnaseq_r20140717 (Grabherr et al., 2011)) and were clustered with OrthoMCL (version 2.0.9 (Li et al., 2003)) including a set of presumed arthropod single-copy genes retrieved from `http://busco.ezlab.org/` (version 1). OrthoMCL clusters were selected as possible targets if they had no more than one sequence per transcriptome sample, and contained a BUSCO database entry. Their alignments were inspected visually. If there were no major disagreements or large indel polymorphisms, consensus sequences were generated. Consensus sequences were then ordered by length. Starting with the second longest, every other one was trimmed to half of its length from the 3′-end until their cumulated length was 1 Mbp. No hits were found when these sequences were compared using BLSTn to a list of known

Table 6: **Individuals sampled for RNA sequencing.**

Individual	Karyotype	Tissue	Clean pairs	Sex	Sample No
14#Y1.5	XY	testes	46 253 164	male	85
14#Y1.5	XY	leg	40 750 427	male	86
14#O1.5	X0	leg	42 596 157	female	87
14#O1.5	X0	ovary	38 887 484	female	88
14#T6	hybrid	brain	43 146 605	male	89
14#T6	hybrid	testes	39 590 463	male	90

repetitive sequences from *P. pedestris* which had previously been generated using the software RepeatExplorer (see Chapter 3). Baits of 120 bp length each were then designed by MYcroarray (Ann Arbor, Michigan) who also generated sequencing libraries, carried out sequence capture, and organised sequencing.

5.2.4 DNA samples and extraction

Samples for DNA sequencing were taken in the field as whole legs, which were stored by immersion in individual plastic 2 ml screw-top tubes (VWA) of pure ethanol. DNA was extracted from femur muscle tissue, which was dissected out of the cuticle in the laboratory, washed twice for 5 min in demineralised water on ice and was then processed using a Qiagen Gentra HMW kit following the manufacturer's instructions. DNA was quantified using a NanoDrop spectrophotometer and a Qubit photometer. The 192 samples sequenced by MYcroarray contained DNA from 393 grasshoppers. While separate sequencing of individual samples is necessary for the analysis of linkage disequilibrium and heterozygosity, sequencing of pooled samples is more cost-efficient when determining genotype frequencies. 138 samples contained single-individual DNA only, 54 samples were pools of DNA from up to five individual grasshoppers.

When collecting samples along transects, priorities were: (1) to sample multiple transects (to discern random fluctuations from the signal of hybridisation), (2) to cover long distances per transect (so as not to miss polymorphism clines), (3) to take samples densely-spaced (to capture gradual changes in frequency), and (4) to collect sizeable samples (to reduce sampling error). In practice, there

are biological, logistical, and economical constrains: the species habitat may not allow long transects, desirable sampling points can be inaccessible, and the more samples, the higher the cost, in particular for single-individual samples. Samples were taken from five transects (number of samples taken and total length are given in parentheses):

- Authon (two, 2.9 km),
- Barles (25, 2.46 km),
- Les Tomples (17 densely-spaced and one remote, 6.3 km),
- La Blanche (seven, 16.6 km making up the longest transect), and
- Lac d'Allos (five, 2.6 km)

While the transects were chosen in order to sample across the transition between X0 and neo-XY, it turned out that all individuals from the hardly accessible Lac d'Allos site were neo-XY. They may be seen as a negative control. Samples were also obtained from the mountains La Bigue and Le Blayeul. Finally, there are samples of an X0 sire (from the northern end of the La Blanche transect), a neo-XY dam (from Le Bayeul) and three of their offspring. The samples from the Les Tomples transect and from the ends of all other transects were sequenced individually. Most other samples where pooled.

The sampling locations are shown in Figure 18, which has a dot for every collection point) a detailed list of samples can be found in Tab. 9 in Appendix C. This includes a column indicating the number of individuals contributing to each sample and the coordinates of each sampling location.

5.2.5 Assembly, filtering, and variant calling

After trimming 3'-ends to a phred quality of at least 30, paired reads were assembled using Platanus (version 1.2.4 http://platanus.bio.titech.ac.jp/). The contigs obtained were aligned using BLASTn to the target sequences and contigs which overlapped others by 10 bp or more were excluded to avoid potential paralogues. Contigs were removed, too, if they showed sequence

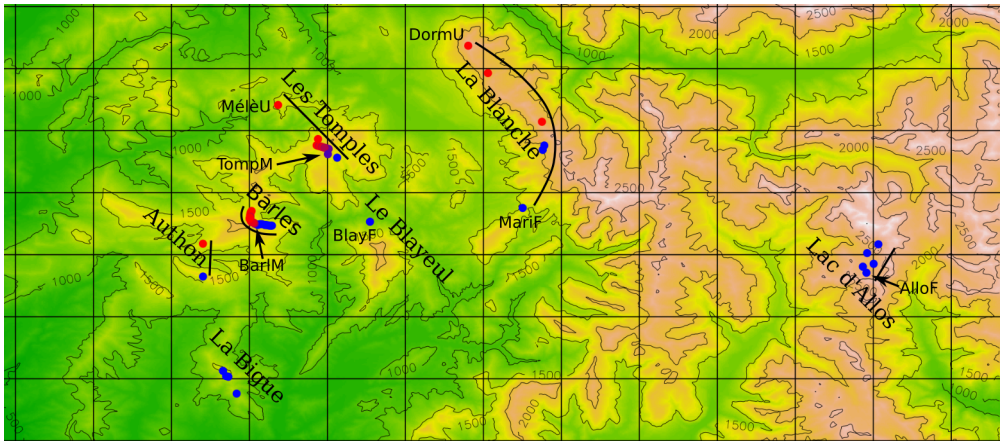

Figure 18: **Sampling sites.** Blue dots label XY (fused) populations, red dots neo-XY (un-fused). The names of transects and mountains are shown in serif font, those of samples used for the isolation-by-distance analysis in sans (see Tab. 7).

similarity to genomic repeats as identified by BLASTn to *P. pedestris* RE clusters identified in Chapter 3. After attempts to call polymorphisms had failed, presumably because of high variability in non-coding regions, STAR (version 2.5.3a) was used to map transcriptome reads to the *de novo* assembly and to determine the positions of exons. They were stored in a BED file which was later used to constrain variant calling to such regions. Mapping to the assembly was carried out using bowtie2 individually for each sample. Variants were called using GAKT (version 3.7 McKenna et al. (2010)) following the best practices for germ line variant detection (DePristo et al., 2011).

5.2.6 Heterozygosity filtering

The subset of 138 single-individual samples was used to calculate per-locus allele frequencies and heterozygosity which are shown in Fig. 19. This analysis identified a number of loci whose heterozygosity equalled the frequency of the rare allele which means each occurrence of the rarer allele was in a heterozygous individual. While this is not an unlikely observation for rare alleles *or alleles private to Y-chromosomes*, loci showing this pattern for higher-frequency alleles are potentially located on contigs of collapsed paralogues. Genomic targets

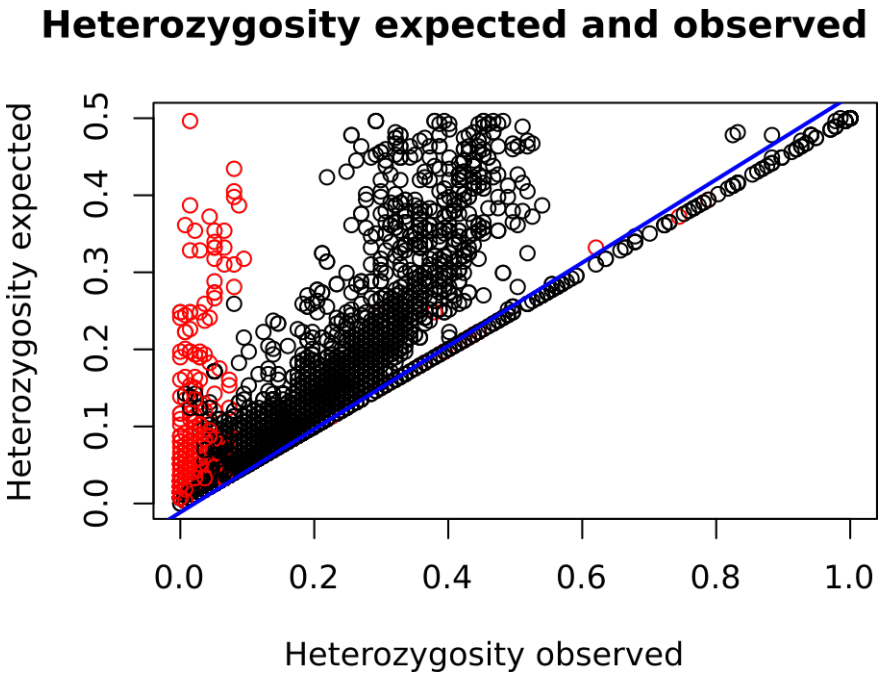

Figure 19: **Per-SNP heterozygosity expected and observed computed from 138 single-individual samples.** Circles below the blue line represent loci filtered out due to high heterozygosity. Red circles indicate low coverage in males and thus putative X-linkage.

which contained such SNP loci were excluded from the analysis. The criterion was $het_{expected} < het_{observed} \times 0.54 - 0.0115$. Given the sampling, this would allow an Y-chromosomal allele present in each neo-XY male to pass the filter. The criterion is reflected by the line in Fig. 19). The remaining set contains 12 548 bi-allelic SNPs.

5.2.7 Diversity measures

Single-individual samples were used to calculate diversity measures and F-statistics. SNP heterozygosity was determined for every individual. Per-SNP F-statistics treating all neo-XY and X0 individuals as populations were calculated following (Weir & Cockerham, 1984) as implemented in the R package pegas (version 0.10 (Paradis, 2010)). Pairwise F_{ST} between sampling sites was computed using the R-package hierfstat (version 0.04-22), which follows Nei (1973).

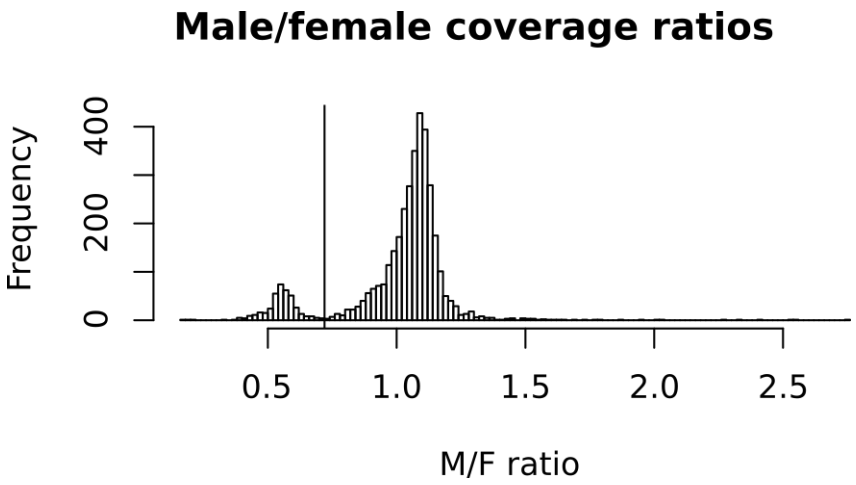

Figure 20: **Per-contig ratios of relative coverage between male and female samples.** A threshold of 0.7 (indicated by the line) was chosen to label putatively X-linked contigs. Note, the peaks are shifted to the right. This is expected because the sexes have different X-to-autosome DNA ratios. (Females have a slightly lower relative amount of autosomal DNA than males.) More generally, the larger the X in an X0 system, the greater the shift expected.

5.2.8 Coverage analysis

Targets with putative X-chromosome linkage were identified through the analysis of differences in mapping depth between male and female samples. This analysis was carried out on contig level. For each contig, a linear model was fitted (coverage explained by sex with 192 observations). The model coefficients obtained were converted into male/female coverage ratios. The histogram summarising the data of all 3704 contigs (shown in Fig. 20) is bimodal, the two peaks presumably representing autosomal and X-chromosomal contigs. A threshold of 0.7 was set by inspection of the histogram below which contigs were labelled as putatively X-linked. All targets containing such contigs were considered putatively X-linked as well.

5.2.9 Cline fitting

Logistic regression models were fitted in R using the function glm to allele frequency data from four transects (Barles, Les Tomples, La Blanche, and

Lac d'Allos). The coefficients of such a model are y-intercept and slope of a line fitted in logit space. These can be converted into cline parameters as $centre = -\frac{slope}{intercept}$ and $width = \frac{4}{slope}$. R's function `glm` is very robust, in particular it will estimate parameters even if the allele frequencies are fixed for zero or one in all populations, and a simple logit transformation would fail. If there is no clear transition in allele frequency within the transect, `glm` will typically fit a shallow cline with its centre displaced to one side or the other of all the sample sites. Consequently, down-stream analyses only used loci with cline-centres fitted inside the transects (a location between the first and last site).

5.2.10 Steep clines

Steep clines were identified from the difference in fitted allele frequencies on either side of the chromosomal transition. The results were compared over four transects, selecting samples so that they covered just over 2 km in each case: La Blanche (the inner four samples), Les Tomples (excluding the far sample), Barles, and Authon. At Authon there were only two samples spanning the zone, so the observed frequencies were used rather than those from a fitted cline. The results obtained are consistent for frequency differences ranging between 30 and 80 %.

5.3 Results

5.3.1 General statistics

Targeted capture was performed on 968 genomic regions (targets) in 192 samples (138 individuals and 54 pools). After de novo assembly, collapsing of contigs and filtering out potentially duplicated regions 2704 contigs remained from 800 targets. These contain 12 458 single-nucleotide polymorphisms. Analysis of the mapping depth of male and female reads aligned to the reference showed lower coverage in males in 78 genomic targets suggesting these are located on the ancestral X-chromosome. These targets contain approximately 7 % of the SNPs detected.

Per-individual proportions of heterozygous loci are shown in Fig. 21. Each

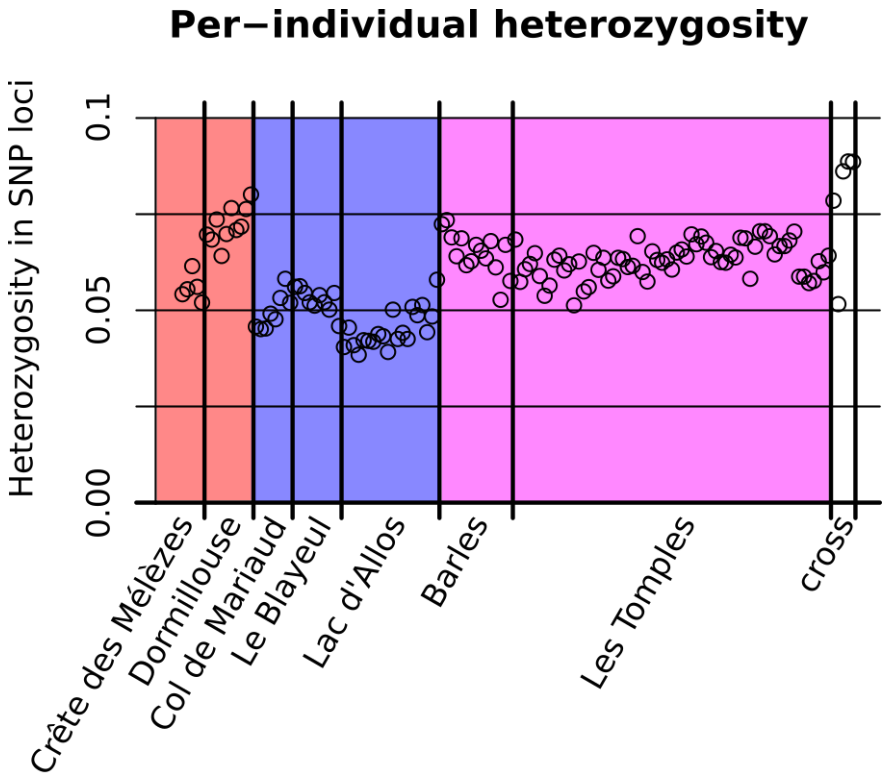

Figure 21: **Fraction of heterozygous loci per individual.** Un-fused locations are shaded in red, fused ones blue, and mixed ones in magenta. A cross is shown on the right-hand side. Individuals are grouped by population, the within-population grouping is arbitrary.

individual contains several hundreds of heterozygous loci. Individuals involved in a cross are shown in the right-hand column. From left to right the five circles represent the sire from Col Bas (near Dormillouse), the dam from Col de Mariaud, and three offspring with overshooting heterozygosity.

F-statistics were computed between the neo-XY and XO populations. Fig. 22 shows the distribution of F_{ST} values of 12 458 SNPs and the relative enrichment of X-linked SNPs in high-F_{ST} categories.

Pair-wise F_{ST}-values were computed between sampling sites (shown in Tab. 7 and in Fig. 23). Interestingly, F_{ST} values are highest for inter-karyotype comparisons, as would be expected if populations with different karyotypes had diverged more. This is supported by a partial Mantel test (999 permutations, Mantel statistic $= 0.7208$, $p = 0.002$). Allos, which is very remote from all other

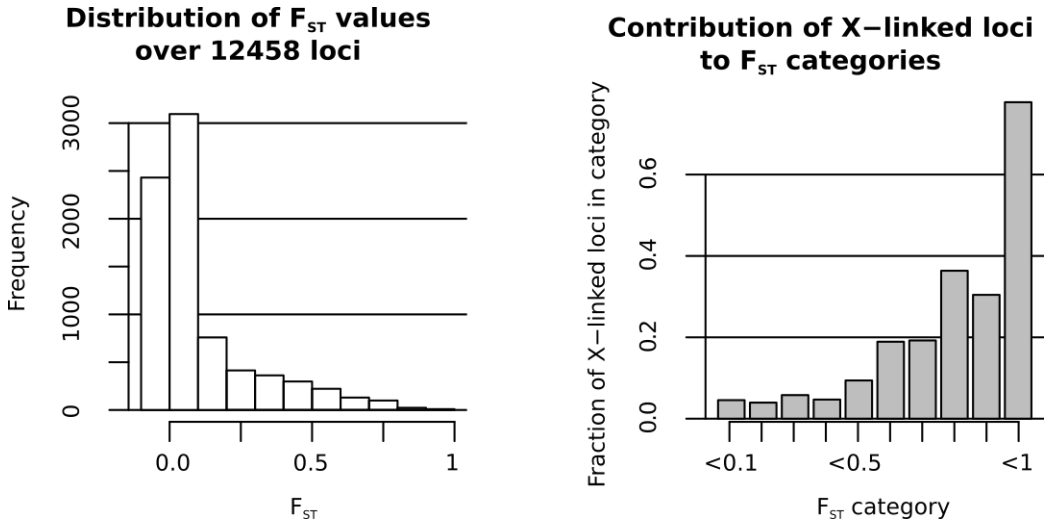

Figure 22: **Left: distribution of per-locus F_{ST} for 12 548 SNPs. Right: Relative contribution of X-linked SNPs the F_{ST} categories.** X-linked loci are strongly over-represented in high-F_{ST} classes.

sites, shows high values in all comparisons the highest being Dormillouse and Crête des Mélèzes with the alternative karyotype.

5.3.2 Steep clines

Following the idea of strong differentiation due to secondary contact, the data were screened for SNPs which show steep clines (allele frequency differences > 0.8) near the X-chromosomal one. The numbers of SNPs with frequency differences $> 80\%$ are shown in Tab. 8 together with the numbers of loci coinciding between transects. There was only one single SNP with a frequency difference > 0.8 in all four transects. This SNP is called 'T0254_Contig1_499' (genomic region 'T025', contig name 'Contig1', position 499). A different way of looking at this is that the data contain only one diagnostic marker (a marker which consistently differs between the groups of interest).

5.3.3 All clines

In order to extend the search to shallower clines and ones that might have their centre further away from the chromosomal cline, the full sampling data was

Table 7: **Pair-wise F_{ST} between sampling sites are shown in the lower triangle, geographical distances (in km) in the upper triangle.** See also Fig. 23. The sites are indicated in Fig. 18. The letters appended to the abbreviated location names indicate the karyotypes present: F – fused (neo-XY), U – un-fused (X0), M – mixed

	AlloF	**BarlM**	**BlayF**	**DormU**	**MariF**	**MélèU**	**TompM**
AlloF		40	32	32	22.5	40	36
BarlM	0.162		7	21	17	11	8
BlayF	0.149	0.068		17	10	12	7
DormU	0.282	0.144	0.239		15	14	13
MariF	0.151	0.087	0.067	0.252		18	14
MélèU	0.185	0.106	0.179	0.185	0.211		4.5
TompM	0.113	0.03	0.035	0.081	0.038	0.019	

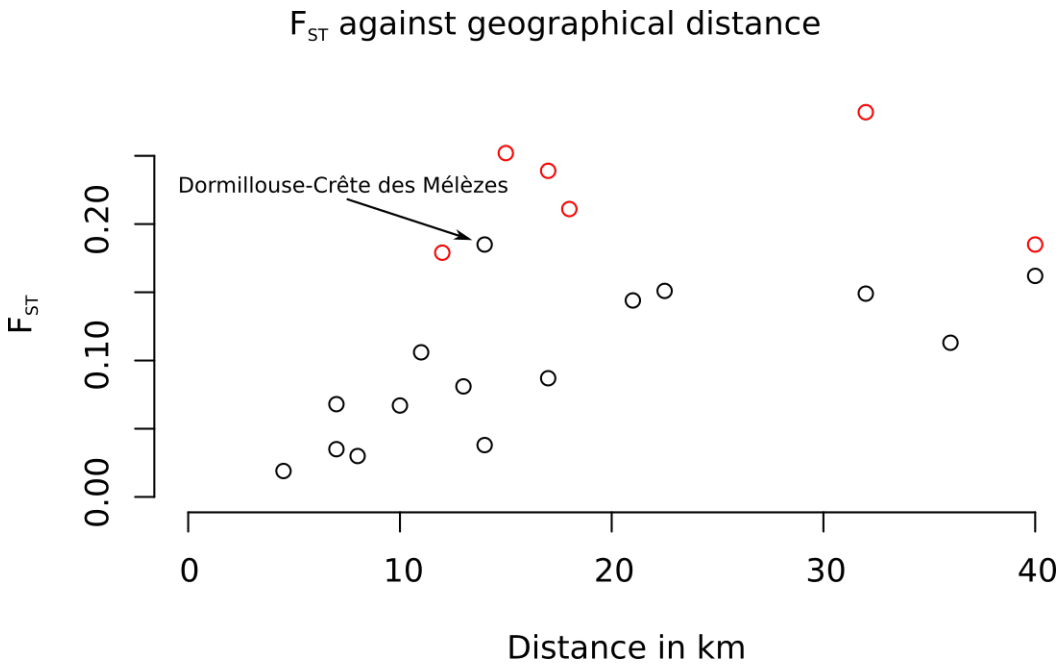

Figure 23: **F_{ST} plotted against geographical distance for seven locations sampled.** Comparisons across the hybrid zone are highlighted in red.

considered from Barles, Les Tomples, La Blanche, and Lac d'Allos. For each transect, cline widths and centres were recorded for those clines whose fitted centres were inside the sample range of that transect. The fraction of loci with

Table 8: **SNPs with allele frequency differences greater then 80%.** Steep clines per-populations are shown on the diagonal. Pair-wise concordance is indicated in the upper triangle. Very few steep clines agree between transects.

	Authon	**Barles**	**Les Tomples**	**La Blanche**
Authon	12	2	1	2
Barles		13	1	7
Les Tomples			15	1
La Blanche				22

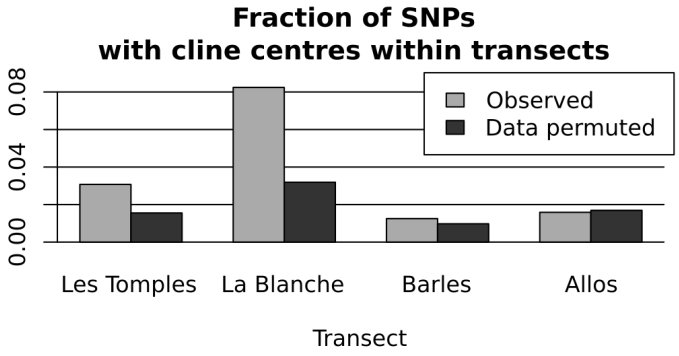

Figure 24: **Fraction of SNPs whose cline centres fall within a transect.**

cline-centres within the transect varies among localities (see light bars in Fig. 24), it seems to correlate with transect length: the longer a transect the more clines were detected with centres within that transect.

The relationship between cline widths and cline centres in four transects is shown in Fig. 25. The SNP T0254_Contig1_499 is highlighted by a cross. It always shows the narrowest cline (except in Allos where the transect does not cross the chromosomal cline). The colours indicate co-occurence in several transects of clines narrower than 5 km (the width expected of a neutral cline after $t = 10000$ years of contact given the species' dispersal rate of $\sigma = 20$ m following (Barton & Gale, 1993) is $2.51\sigma\sqrt{t} = 5020$ m). Highlighted in red are loci which have narrow clines in two transects, blue indicates occurrence in three transects. There is no narrow cline that occurs in all four transects.

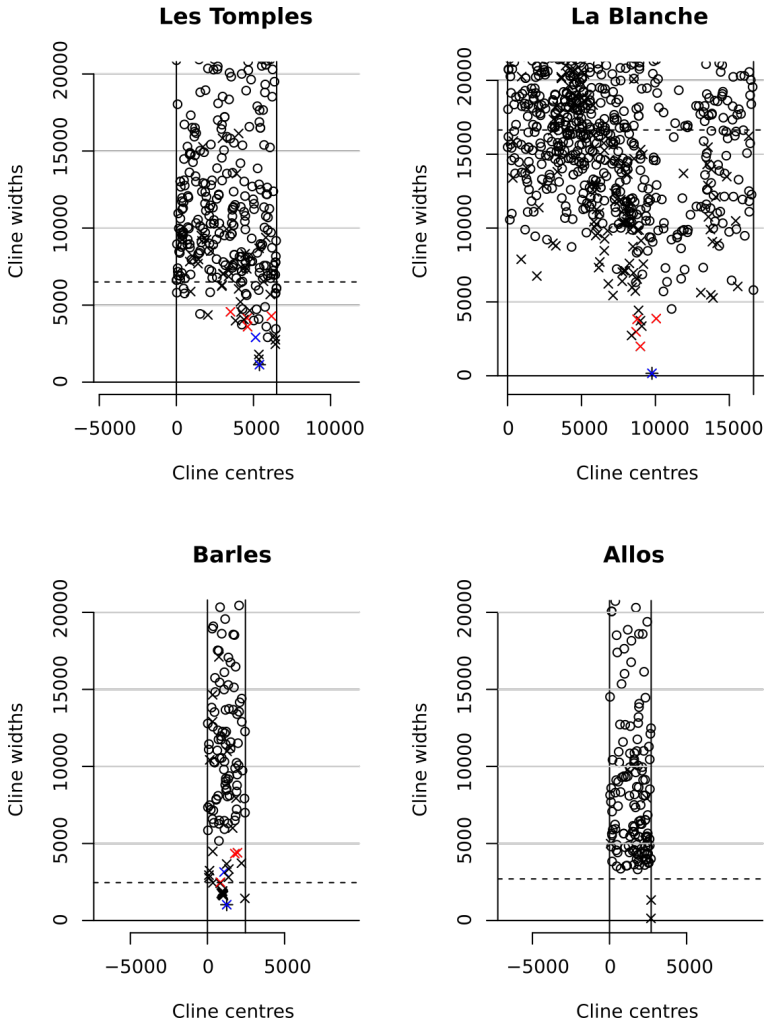

Figure 25: **Widths, centres, and concordance of clines fitted to SNP data.** For comparison to cline widths, the length of each transect is indicated on the y-axis as a dashed line. T0254_Contig1_499, the SNP which is highly correlated with the chromosomal fusion is indicated by a plus sign. Clines with widths less than 5 km are highlighted in red if they occur in at least two transects and in blue if they are shared by three transects. No cline narrower then 5 km is shared by all transects. X-linked SNPs are shown as Xes. All plots shown have the same aspect ratio.

X-linked loci are shown as Xes. These loci show some of the narrowest clines in each subplot.

The distribution of cline widths per transect is shown as histograms in Fig. 26.

The La Blanche transect, which extends far into both pure populations, shows a higher median cline width than any other transect. This might be because neutral introgression has caused wide clines which can only be detected in long transects. Alternatively, in a long transect one might pick up signals of isolation-by-distance which may look like the signal of introgression. To test this, the allele frequency data were permuted randomly between the sampling sites of each transect, and clines were fitted to the permuted data. The results of 1000 permutations are shown as a curve in Fig. 26. While the distributions of permuted and observed data are similar in the short transects (Lac d'Allos and Barles), very few clines fitted to the permuted data have centres within the long clines (Les Tomples and La Blanche), suggesting that shallow clines can only be detected in longer transects.

5.4 Discussion

The targeted capture sequencing carried out on 192 samples from several transects crossing the *Podisma pedestris* hybrid zone gives insight into the differentiation across the hybrid zone and between populations.

5.4.1 Concordance of clines between transects

The shape and position of clines forming in a zone of secondary contact will depend on whether selection is acting and how much. All else being equal the width of a cline of one underdominant locus will, at equilibrium, depend on the balance between dispersal and selection, while neutral clines will become wider over time (Barton & Gale, 1993). Both kinds will have their centre at the same point.

It may be helpful to ask, what happens if such a system is disturbed? While neutral clines may be disrupted and mixed by genetic drift, for instance during a period of habitat fragmentation, a cline maintained by selection is expected to re-establish similar to a Gömböc[2] finding its stable point.

[2]Gömböc is the name of a spherical Hungarian sausage. The word is also used for the mono-monostasic solid discovered by Varkonyi & Domokos (2006), which is a homogeneous body which, when placed on an even surface, will always roll back into the same position.

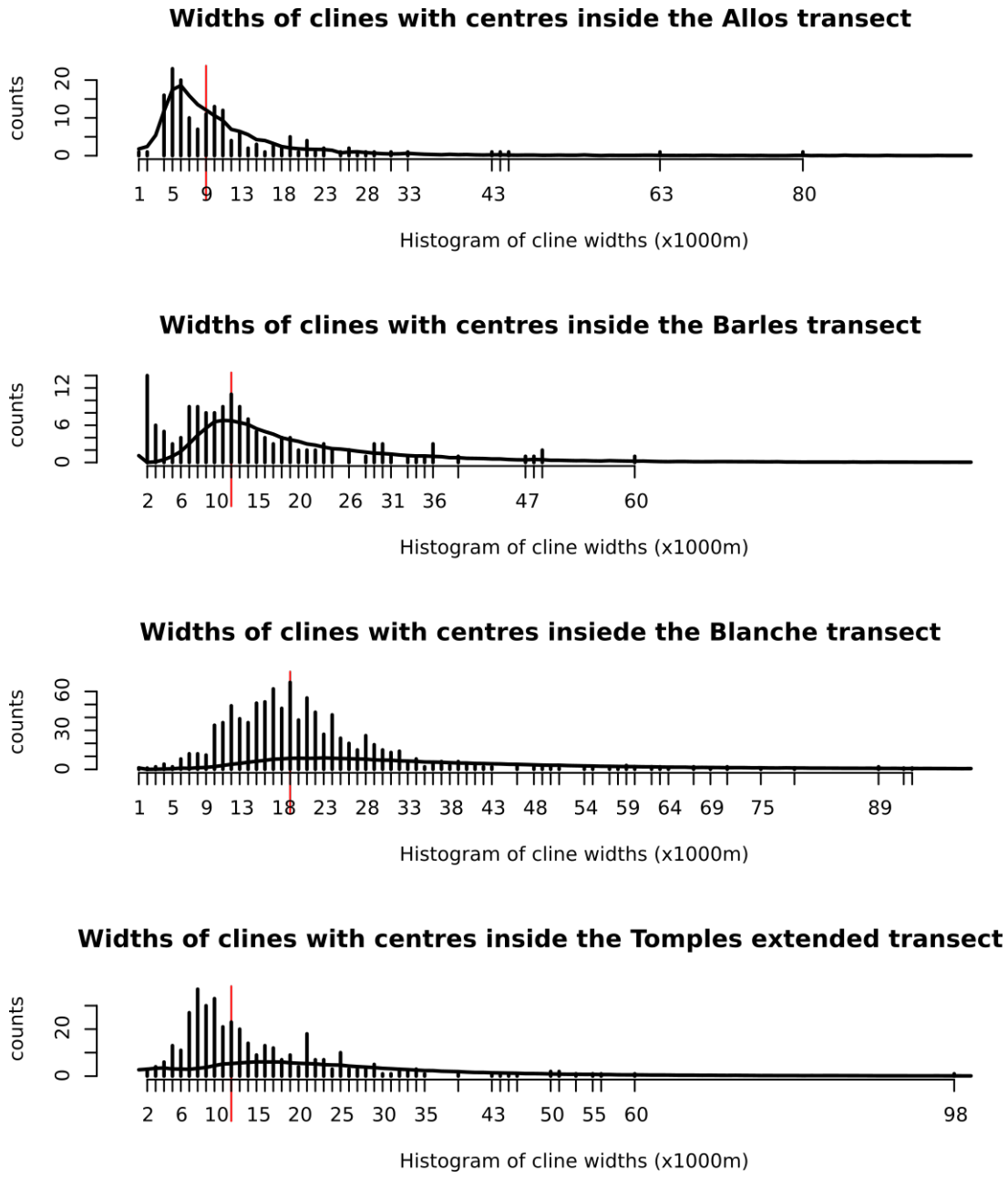

Figure 26: **Distribution of cline widths in each transect.** The distributions observed are shown as histograms. Median cline widths are indicated by red lines. The curves indicate the distribution of randomised data (1000 permutations).

This is why the comparison of clines between transects can provide stronger evidence of selection than merely looking for steep clines. If selection is acting the same way in different transects, selected clines should show the same pattern in all transects. Of all the clines identified, there are only two which are narrower than 5 km (the expected width assuming neutral diffusion with a standard deviation of 20 m in a generation and 10 000 generations since the two taxa met at the end of the last ice age). These are T0254_Contig1_499 and T0128_singlet_16848505_0618. Both are X-linked, in each of the transects their centres lie within 300 m of each other, and their widths are similar in the Tomples and Barles transects (The Blanche transect is ignored here because it was sampled too sparsely for accurate estimates of width and location).

Very few of the clines identified show similar patterns in different transects. This may be due to Holocene disturbance of the part of the range in which Les Tomples and Barles are located. They are on an isolated mountain block, which mostly does not rise higher than 1800 m. During the Holocene climatic optimum the warmer temperatures may have driven *P. pedestris* further up the mountains and will have considerably reduced the area hospitable to the species in this area. Once the climate cooled again, *P. pedestris* would have re-colonised lower areas and drift-related processes may have caused the spread of recombined allele combinations (Nichols & Hewitt, 1986). This disturbance may have removed (or reduced) the expected signal of neutral introgression after secondary contact, whereas the clines found to be concordant between transects are good candidates for loci under selection.

5.4.2 Genetic architecture of the hybrid zone

Following the calculation that the X-chromosomal cline, which is 800 m wide, could be maintained by selection of only 1 % against chromosomal heterozygotes (Barton & Hewitt, 1981a), the discovery of strongly reduced viability in hybrids produced by crossing (of the order of 50 % reduction (Barton, 1980)) appeared to be paradoxical. To resolve the paradox Barton & Hewitt (1981b), postulated that this selection was not concentrated on the chromosomal difference, but that hundreds of differentiated loci with comparable weak selection might differ

between the two sides of the zone.

One potential problem with this explanation is that clines of many weakly selected loci can attract one another causing strong selection against early-generation hybrids and causing a barrier to gene flow. This depends on the balance between recombination and selection acting. If selection is stronger, then neutral loci would show steep clines because they are in linkage disequilibrium with selected ones. If however recombination has a stronger influence than selection, clines will not show a stepped pattern. There would still be strong selection against inter-population hybrids, but the cline would not necessarily act as a strong barrier to neutral gene flow (Barton, 1983). The stepped pattern of a genomic cline similar to the one found in the *Bombina* hybrid zone (Szymura & Barton, 1991) was not detected here.

Out of 12 458 SNP loci from 800 genomic regions only a single one shows a cline as steep as the X-chromosomal one. In fact, this SNP (T0254_Contig1_499) is likely located on the ancestral part of the X chromosome and its allelic state is strongly correlated with the presence of the X-chromosomal fusion. Most other SNPs showing comparatively steep clines are X-chromosomal too (see Fig. 25 where X-chromosomal loci are shown as Xes). These steep clines occur within few a kilometres with the X cline. None of these loci are diagnostic across all transects. Rather than locally selected alleles, they may represent loci linked to selected ones, which only become visible because of geographically limited alleles. Notably the Lac d'Allos transect, which does not cross the centre of the X chromosome cline, does not have any SNP clines narrower than the transect is long (see bottom right plot in Fig. 25) except for two X-chromosomal SNPs with centres fitted just inside the boundary.

The longest transect (which follows the mountain ridge of La Blanche) contains the highest number of cline centres, presumably because long transects are required in order to detect shallow clines. The distribution of cline widths (shown in Fig. 26) has a median of 19 km for La Blanche which is much wider than the neutral expectation of 5 km. How can this discrepancy be explained?

First, there might be more opportunity in a long transect to pick up the effect of drift: populations could have randomly drifted to different allele frequencies. The permutation analysis suggests that this explanation is unlikely: there would

be far fewer random clines with centres within the La Blanche transect. A second possible explanation is that *P. pedestris* might disperse further than Barton and Hewitt's (1982) estimate; although a very similar estimate has been obtained using a different method (Mason et al., 1995). Yet another possibility is that the pattern observed is not the result of diffusion-like introgression between two parapatric populations. Rather, long-distance dispersal during first colonisation of the habitat may have resulted in a patchy distribution of genotypes in space (Nichols & Hewitt, 1994). While differences in neutral alleles would have been smoothed out, alleles showing underdominance would have been pushed into clines. Clines at different loci might then be displaced from each other by selection to minimise the opportunity for contact between the two forms, a phenomenon suggested for the *Chorthippus parallelus* hybrid zone, discussed by (Virdee & Hewitt, 1994).

5.4.3 The role of the sex chromosomes

In the transects sampled here, the only SNPs showing steep clines are located on the X chromosome. It can be concluded that selection is acting on the (ancestral) X chromosome. Is the X chromosome special?

Sex chromosomes are unusual genomic regions expected to behave differently to the rest of the genome. X (and Z) chromosomes are hemizygous in the heterogametic sex, exposing recessive (and partially recessive) alleles to selection. Y and W chromosomes, if present at all, are always hemizygous. Sex chromosomes may have different (smaller) effective population sizes than autosomes. Their recombination rates can be lower as well as higher than those of autosomes. X chromosomes can show lower recombination rates than autosomes, because they can recombine in females only whereas autosomes may recombine in both sexes, but in species of *Drosophila*, which lack recombination in males, X chromosomes (having a lower effective population size than autosomes) show higher effective recombination rates. Charlesworth et al. (1987) showed in a theoretical study that beneficial alleles are expected to become fixed more rapidly on sex chromosomes if they are partially recessive and that the fixation of underdominant rearrangements is expected to be faster with sex

chromosome linkage, possibly causing higher rates of adaptive evolution. But the evidence for faster evolution of X linked genes is mixed, which may partly be because theory focussed on adaptive changes whereas tests often do not discriminate between synonymous and non-synonymous changes (Meisel & Connallon, 2013).

There are two empirical rules which point out the importance of X-chromosomes to speciation, Haldane's rule and the large-X effect. Haldane's rule points out that if in a cross only one sex suffers a reduction in fitness, it tends to be the heterogametic one (Haldane, 1922). The large-X effect is based on the observation that in Drosophila, species incompatibilities are often linked to the X (Coyne, 1992). While there is no differential hybrid unfitness between the sexes in *P. pedestris*, the X-clines detected do agree with a potential large-X effect. Lasne et al. (2017) recently pointed out the potential importance of the large-X effect in local adaptation. The X-linked loci identified may be under selection but without close linkage to the fusion.

There is another consideration. It might be just more likely to detect selected sites on the X-chromosome if recombination is reduced relative to autosomes in *P. pedestis*, potentially causing linkage disequilibrium to extend further than on autosomes. Where linkage disequilibrium extends further from a selected site, it is more likely to overlap with a site targeted in the sequence capture experiment.

5.4.4 Mating system

A congeneric species of *P. pedestris*, *P. sapporensis*, has X0 and XY populations too. Warchałowska-Śliwa et al. (2008) have shown that offspring of inter-population crosses in this system show a high frequency of parthenogenetic individuals. These grasshopper's somatic tissues are mosaics of haploid and diploid cells suggesting a form of parthenogenesis through heterozygous eggs with subsequent doubling of the chromosome set.

This form of reproduction would reduce gene flow between populations and might have the potential to cause steep genomic clines. Hewitt et al. (1987) found that female *P. pedestris* individuals which had mated in the hybrid zone to have

preferentially homogamic offspring, that is, few were heterozygous for the sex chromosome system. A follow-up study looked at females mated sequentially with one male of each population. There were two main observations: The sperm of the first mate is more likely to fertilise and there was again a bias towards homogamy, suggesting a prezygotic postmating mechanism of isolation. There also was a bias in sex ratio, more offspring being female. However, Hewitt et al. (1989) argue that parthenogenesis seems unlikely to have had a major influence on their observations.

The analysis of per-individual heterozygosity based on the data generated here did not find any evidence for grasshoppers generated through parthenogenesis. These would be expected to show virtually no heterozygous sites causing high variability in the levels of heterozygosity observed per population. This was not the case (shown in Fig. 21). The progeny of the cross set up (Fig. 21, right-hand margin) show high heterozygosity too. However, these observations certainly do not exclude the possibility of rare parthenogenesis and it would be worthwhile to explore the potential effects on hybrid zone dynamics.

5.4.5 The power of hybrid zone analysis

This chapter demonstrates the power of hybrid zone analysis. Comparatively steep clines may occur at random through neutral processes, but the comparative analysis of multiple transects shows which loci are under selection. Two SNPs under selection were detected here, with cline widths of approximately 1000 m and 3300 m. This corresponds to selection coefficients of 0.3 % and 0.03 %, which would have been impossible to pick up in a crossing experiment.

5.5 Conclusion

The genetic architecture of the *P. pedestris* hybrid zone was investigated using SNP data from hundreds of loci across the genome. Fitting sigmoidal clines to genotyping data obtained from multiple transects crossing the zone, only two loci under selection could be identified, both X-chromosomal. The lack of any autosomal loci showing signatures of selection prompts further research into the nature of the selection against hybrids observed in the zone. The results

show impressively how hybrid zones can be exploited to detect weak selection in action.

6 General Discussion

This thesis presents the first study of the *Podisma pedestris* hybrid zone which makes use of high-throughput sequencing data. After summarising what we have learned about the hybrid zone, this discussion will re-examine and develop the interpretation of the molecular differentiation between the two populations, will attempt to explain the origin and spread of the sex-chromosomal fusion, will touch on the relevance of theories of speciation by reinforcement, and will finally suggest avenues for future research on *P. pedestris*.

6.1 What have we learned

Chapter 2 explored the potential effect on the *P. pedestris* hybrid zone of the 'inexorable spread' dynamics, modelled by Veltsos et al. (2008). This study suggests that a neo-sex chromosome system could replace the ancestral one in a process fuelled by sexually antagonistic selection. This form of selection is one possible explanation for the spread of the fused X and the neo-Y chromosomes into the genomic background of the X0 population.

It seems plausible that the requirements for inexorable spread are fulfilled: there is reduced recombination between the sex chromosomes and sexually antagonistic alleles present in these reduced-recombination areas. But there is no evidence that the phenomenon has had an effect on the *P. pedestris* hybrid zone from the analysis of zone-position in Chapter 2. In particular, the distribution of the sex chromosome coincides closely with a major watershed, which would have been where the populations met after the last glacial maximum.

In addition to the sex-chromosome cline, an second possible source of information about the history of range expansion would be a study of the geographical distribution of mitochondrial DNA haplotypes. However, the only previous attempt to carry out such a study found that this information was obscured by the exceptionally high number of mitochondrial sequences that have been translocated into the *P. pedestris* genome, known as Numts (Bensasson et al., 2000).

Chapter 3 addressed the challenge of separating nuclear and mitochondrial genotypes in a study investigating the relative enrichment of mitochondrial sequences and Numts in long and short-read sequencing data. The chapter provides the complete mitochondrial sequences for six *P. pedestris* individuals and it gives two estimates of the genomic proportion of Numts in the species' genome. The data generated will be of use to further studies of *P. pedestris* and grasshoppers in general. The enrichment found for Numts sequences in long-read data and enrichment for actual mitochondrial sequences in short-read data will allow future studies of mitchondrial phylogeography, and the dynamics of Numt insertion into the genome.

In Chapter 4, a recently developed approach to reconstruct population histories (Lohse et al., 2011) was used to explore the *P. pedestris* demography. Instead of making use of genetic data from multiple individuals from many locations, it exploits the large amount of information encoded in just two diploid transcriptomes. The analysis finds the split between the populations happened several glacial cycles ago, and also raises the question of whether there has been more gene flow between the populations than previously thought.

Chapter 5 revisits the classical analysis of clines in a hybrid zone – but instead of surveying a few tens of loci, it exploits targeted sequence-capture to analyse over 800 loci and some 10 000 polymorphic sites. These SNP clines are analysed and screened for patterns of selection. Given the power of the approach, it is surprising that only two SNPs under selection could be identified. Both are located on the (ancestral) X chromosome. This prompts the search for other possible explanations for the strong selection observed against hybrids.

6.2 Differentiation

How different are the *P. pedestris* populations? Our view on *P. pedestris* is certainly centred on the populations differing in their sex chromosomes. But clearly this flightless grasshopper does not disperse very far (Barton & Hewitt, 1982) and considerable structure within the populations might be expected. Is

differentiation between the populations generally higher than within? Fig. 23 on page 78 shows pair-wise F_{ST} plotted against geographical distance. While there is an increase in F_{ST} with increasing distance, comparisons across the hybrid zone (X0–XY, highlighted in red) tend to show higher F_{ST} values, reflecting two somehow differentiated populations coming together after the last glacial maximum.

But this picture is not perfect, the comparison Crête des Mélèzes-Dormillouse (both X0) shows a rather high FST value, on a comparatively short geographical distance. This may be a sign of the comparatively low genetic diversity of the Crête des Mélèzes sample (reflected by its comparatively low heterozygosity, see Fig. 21 on page 76), which could be explained by its post-glacial history. The site is located just above 1600 m and may have been extirpated during the Holocene climatic optimum because of the hotter climate. The current grasshoppers are likely the descendants of a re-colonisation from an area nearer to the hybrid zone on Les Tomples. This example shows that, while there are good reasons to assume that the populations differ in the genetic diversity, the sex chromosomes alone are just one locus under selection, which have distinct histories and can show different patterns from the rest of the genome. This phenomenon will be of less importance in areas with higher-elevation peaks where there was not much disturbance during the Holocene. But such disturbance of neutral or nearly neutral diversity can be helpful for detecting loci under selection, whose clines are expected to re-emerge after disturbance possibly showing a contrasting pattern.

6.3 Origin of the fused population

It is generally assumed that chromosomal rearrangements are rare and, if they are selected against when heterozygous, they will additionally have a low probability of becoming fixed in a population. But is it impossible that the X-to-autosome fusion happened multiple times in *P. pedestris*? Charlesworth & Charlesworth (1980) show in a theoretical study that the presence of sexually antagonistic selection (more precisely a selectively maintained allele frequency difference between the sexes at a locus) will favour the fusion of such a locus

to a sex chromosome. It is far more likely that a fusion happens to the Y chromosome, causing stronger correlation between sex and allelic state (because Ys are confined to males whereas Xes occur in both sexes). The fusion which happened in *P. pedestris* is an autosome-to-X fusion, but one might argue it had the same effect. By creating a neo-Y, there is now a male restricted element (the region of reduced recombination, discussed in Chapter 2). It is unclear how the area of reduced recombination was first established. It might be the result of selection on the neo-sex chromosomes to keep male-beneficial and female-deleterious alleles confined to males. Alternatively, there may have been an area of reduced recombination on a pair of autosomes which encouraged the fixation of the chromosomal fusion. But this would then raise the question of how the area of reduced recombination first evolved.

The grasshopper *Podisma sapporensis*, too, has X0 and neo-XY individuals. A mitochondrial tree does not show karyotypes as monophyletic, which caused Tatsuta et al. (2006) and Kowalczyk et al. (2009) to ask whether the fusion arose multiple times. It seems not unlikely however that this pattern was caused simply by gene flow between populations.

6.4 Reinforcement

The possibility of enhanced prezygotic isolation evolving in sympatry or parapatry has intrigued generations of workers. Following Dobzhansky's (1937) report of stronger reproductive isolation between parapatric *Drosophila* populations than between distantly sampled ones, the idea that reproductive isolation becomes established by reinforcement became very popular and for some, it became the main driving force of speciation, even in absence of strong experimental evidence (Coyne & Orr, 2004). Butlin (1987) pointed out that many cases cited as examples for reinforcement were equivocal. He also described a related phenomenon that should be considered separately from reinforcement, suggesting the term 'reproductive character displacement'. Reproductive character displacement happens when diverged forms produce completely unfit hybrids so that no gene flow is possible. In such a case selection might still favour prezygotic isolation to reduce the wastage of gametes. Recent theoretical

studies are more positive about the possibility of reinforcement if female choice is involved (Liou & Price, 1994; Kelly & Noor, 1996).

P. pedestris shows a trend towards homogamy: inseminated females taken from the hybrid zone were more likely to have offspring homozygous for their sex chromosomes (Hewitt et al., 1987). Because there is no evidence for assortative mating in *P. pedestris*, they suggested that a postmating prezygotic mechanism might be acting. Is it possible that this has evolved in parapatry as the result of gametes wasted to unfit progeny? First of all it is questionable what advantage this would possibly provide and to whom. Postmating isolation would neither prevent sperm being wasted nor energy being invested into courtship and mating. One could argue it might benefit the populations by preventing them from being invaded by fitness-reducing allele combinations. But it seems unlikely there would be a benefit on an individual-basis.

Irrespective of the mechanism, had prezygotic isolation evolved as a result of secondary contact, then this isolation would be expected to be weaker in crosses between individuals taken far away from the hybrid zone. If this was true, there would be an inverse cline (Caisse & Antonovics, 1978) for homogamy. This could be tested, but would require large quantities of offspring to be screened as Hewitt et al. (1989) point out. Further, the importance for *P. pedestris* as a species may be small because of the 'swamping' effect of gene flow towards the hybrid zone. There is no obvious reason why alleles causing assortative mating would have spread away from the hybrid zone into regions where they have not benefit (because there is no potential the production of hybrids of low fitness). However if reinforcement within the zone were to increase prezygotic isolation, this might eventually reduce gene flow between the populations to an extent that they can diverge further.

6.5 Next steps

This study was the first to tackle the *P. pedestris* hybrid zone with high-throughput sequencing data. It found that the distribution of sex chromosome systems agrees with the ranges the two populations likely colonised after the last glacial maximum. A relatively ancient date was put on the time of the split

between the populations (400 000 years, corresponding to three glacial cycles), and surprisingly few loci under selection were found. What are the next steps?

What are the actual targets of selection? Two SNPs identified here are likely subject to linked selection. What are the loci on which selection acts directly? A genome assembly and, similarly important, a linkage map could help to make sense of these results.

Are the sex chromosomes differentiated? A genome assembly might also give insight into whether the neo-sex chromosomes are differentiated and how. What is the distribution of coverage differences between male and female data like over all contigs? Mapping of paired short or long reads to the assembly might even give a clue as to whether the neo-X and neo-Y chromosomes differ by an inversion.

What is the fine-scale colonisation history of *P. pedestris?* This thesis focuses mainly on samples from the area near Seyne-les-Alpes. Fig. 5 on page 16 shows the neo-XY population inhabits several lower-order watersheds. How differentiated are populations of the same karyotype from different watersheds? Are there clines between the populations?

What causes low hybrid fitness? Given the power of the targeted capture experiment, it would be surprising if the selection against hybrids were due to many loci of small effect. A new cause of low fitness is needed. *Wolbachia* has been implicated in hybrid unfitness in *Chorthippus parallelus* (Martínez-Rodríguez et al., 2013) and endosymbionts might be good candidates to check for.

References

Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, & Gnirke A (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, **12**:R18. doi: 10.1186/gb-2011-12-2-r18.

Avise JC, Arnold J, Ball RM, Bermingham E, Lamb T, Neigel JE, Reeb CA, & Saunders NC (1987). Intraspecific Phylogeography: The Mitochondrial DNA Bridge Between Population Genetics and Systematics. *Annual Review of Ecology and Systematics*, **18**:489–522.

Ballard JWO & Rand DM (2005). The Population Biology of Mitochondrial DNA and Its Phylogenetic Implications. *Annual Review of Ecology, Evolution, and Systematics*, **36**:621–642. doi:10.1146/annurev.ecolsys.36.091704.175513.

Barton NH (1979a). Gene flow past a cline. *Heredity*, **43**:333–339. doi:10.1038/hdy.1979.86.

Barton NH (1979b). The dynamics of hybrid zones. *Heredity*, **43**:341–359. doi: 10.1038/hdy.1979.87.

Barton NH (1980). The fitness of hybrids between two chromosomal races of the grasshopper *Podisma pedestris*. *Heredity*, **45**:47–59. doi:10.1038/hdy.1980.49.

Barton NH (1983). Multilocus Clines. *Evolution*, **37**:454–471. doi:10.2307/2408260.

Barton NH & Gale KS (1993). Genetic Analysis of Hybrid Zones. In Harrison RG, ed., *Hybrid Zones and the Evolutionary Process*, chap. 2, p. 364. Oxford University Press, New York.

Barton NH & Hewitt GM (1981a). A Chromosomal Cline in the Grasshopper *Podisma pedestris*. *Evolution*, **35**:1008–1018. doi:10.2307/2407871.

Barton NH & Hewitt GM (1981b). The genetic basis of hybrid inviability in the grasshopper *Podisma pedestris*. *Heredity*, **47**:367–383. doi:10.1038/hdy.1981.98.

Barton NH & Hewitt GM (1982). A measurement of dispersal in the grasshopper *Podisma pedestris* (Orthoptera: Acrididae). *Heredity*, **48**:237–249. doi:10.1038/hdy.1982.29.

Barton NH & Hewitt GM (1985). Analysis of Hybrid Zones. *Annual Review of Ecology and Systematics*, **16**:113–148. doi:10.2307/2097045.

Bensasson D, Zhang DX, Hartl DL, & Hewitt GM (2001). Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends in Ecology & Evolution*, **16**:314–321. doi:10.1016/S0169-5347(01)02151-6.

Bensasson D, Zhang DX, & Hewitt GM (2000). Frequent Assimilation of Mitochondrial DNA by Grasshopper Nuclear Genomes. *Molecular Biology and Evolution*, **17**:406–415. doi:10.1093/oxfordjournals.molbev.a026320.

Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G, Pütz J, Middendorf M, & Stadler PF (2013). MITOS: Improved de novo metazoan mitochondrial genome annotation. *Molecular Phylogenetics and Evolution*, **69**:313–319. doi:10.1016/j.ympev.2012.08.023.

Bialozyt R, Ziegenhagen B, & Petit RJ (2006). Contrasting effects of long distance seed dispersal on genetic diversity during range expansion. *Journal of Evolutionary Biology*, **19**:12–20. doi:10.1111/j.1420-9101.2005.00995.x.

Boore JL (1999). Animal mitochondrial genomes. *Nucleic Acids Research*, **27**:1767–1780. doi:10.1093/nar/27.8.1767.

Buggs RJA (2007). Empirical study of hybrid zone movement. *Heredity*, **99**:301–312. doi:10.1038/sj.hdy.6800997.

Butlin R (1987). Speciation by reinforcement. *Trends in Ecology & Evolution*, **2**:8–13. doi:10.1016/0169-5347(87)90193-5.

Caisse M & Antonovics J (1978). Evolution in closely adjacent plant populations. *Heredity*, **40**:371–384. doi:10.1038/hdy.1978.44.

Charlesworth B & Charlesworth D (2000). The degeneration of Y chromosomes. *Philosophical Transactions of the Royal Society of London Series B, Biological sciences*, **355**:1563–72. doi:10.1098/rstb.2000.0717.

Charlesworth B, Coyne JA, & Barton NH (1987). The Relative Rates of Evolution of Sex Chromosomes and Autosomes. *The American Naturalist*, **130**:113–146. doi:10.1086/284701.

Charlesworth D & Charlesworth B (1980). Sex differences in fitness and selection for centric fusions between sex-chromosomes and autosomes. *Genetical Research*, **35**:205–214. doi:10.1017/S0016672300014051.

Cox RM & Calsbeek R (2009). Sexually antagonistic selection, sexual dimorphism, and the resolution of intralocus sexual conflict. *The American Naturalist*, **173**:176–87. doi:10.1086/595841.

Coyne JA (1992). Genetics and speciation. *Nature*, **355**:511–515. doi:10.1038/355511a0.

Coyne JA & Orr HA (2004). *Speciation*. Sinauer, Sunderland, 545 pp.

Dayama G, Emery SB, Kidd JM, & Mills RE (2014). The genomic landscape of polymorphic human nuclear mitochondrial insertions. *Nucleic Acids Research*, **42**:12 640–12 649. doi:10.1093/nar/gku1038.

Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, & Pritchard JK (2009). Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, **25**:3207–3212. doi:10.1093/bioinformatics/btp579.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, & Daly MJ (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**:491–498. doi:10.1038/ng.806.

Dobzhansky T (1937). *Genetics and the origin of species*. Columbia University Press, New York, 364 pp.

Dodsworth S (2015). Genome skimming for next-generation biodiversity analysis. *Trends in Plant Science*, **20**:525–7. doi:10.1016/j.tplants.2015.06.012.

Doležel J, Bartoš J, Voglmayr H, & Greilhuber J (2003). Letter to the editor. *Cytometry*, **51A**:127–128. doi:10.1002/cyto.a.10013.

Du Buy HG & Riley FL (1967). Hybridization between the nuclear and kinetoplast DNA's of Leishmania enriettii and between nuclear and mitochondrial DNA's of mouse liver. *Proceedings of the National Academy of Sciences*, **57**:790–797.

Emms DM & Kelly S (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, **16**:157. doi:10.1186/s13059-015-0721-2.

Erler S, Ferenz HJ, Moritz RFA, & Kaatz HH (2010). Analysis of the mitochondrial genome of *Schistocerca gregaria gregaria* (Orthoptera: Acrididae). *Biological Journal of the Linnean Society*, **99**:296–305. doi:10.1111/j.1095-8312.2009.01365.x.

Felsenstein J (1988). Phylogenies from Molecular Sequences: Inference and Reliability. *Annual Review of Genetics*, **22**:521–565. doi:10.1146/annurev.ge.22.120188.002513.

Foerster K, Coulson T, Sheldon BC, Pemberton JM, Clutton-Brock TH, & Kruuk LEB (2007). Sexually antagonistic genetic variation for fitness in red deer. *Nature*, **447**:1107–1110. doi:10.1038/nature05912.

Garber RC & Yoder OC (1983). Isolation of DNA from filamentous fungi and separation into nuclear, mitochondrial, ribosomal, and plasmid components. *Analytical Biochemistry*, **135**:416–422. doi:10.1016/0003-2697(83)90704-2.

Gellissen G, Bradfield JY, White BN, & Wyatt GR (1983). Mitochondrial DNA sequences in the nuclear genome of a locust. *Nature*, **301**:631–634. doi:10.1038/301631a0.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, & Regev A (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, **29**:644–652. doi:10.1038/nbt.1883.

Gregory TR (2016). Animal Genome Size Database.

Haldane JBS (1922). Sex ratio and unisexual sterility in hybrid animals. *Journal of Genetics*, **12**:101–109. doi:10.1007/BF02983075.

Halliday RB, Barton NH, & Hewitt GM (1983). Electrophoretic analysis of a chromosomal hybrid zone in the grasshopper *Podisma pedestris*. *Biological Journal of the Linnean Society*, **19**:51–62. doi:10.1111/j.1095-8312.1983.tb00776.x.

Halliday RB, Webb SF, & Hewitt GM (1984). Genetic and chromosomal polymorphism in hybridizing populations of the grasshopper *Podisma pedestris*. *Biological Journal of the Linnean Society*, **21**:299–305. doi:10.1111/j.1095-8312.1984.tb00367.x.

Hazkani-Covo E, Zeller RM, & Martin W (2010). Molecular Poltergeists: Mitochondrial DNA Copies (numts) in Sequenced Nuclear Genomes. *PLoS Genetics*, **6**:e1000 834. doi:10.1371/journal.pgen.1000834.

Hebert PDN, Ratnasingham S, & DeWaard JR (2003). Barcoding Animal Life: Cytochrome c Oxidase Subunit 1 Divergences among Closely Related Species. *Proceedings: Biological Sciences*, **270**:S96–S99.

Hewitt GM (1975). A sex-chromosome hybrid zone in the grasshopper *Podisma pedestris* (Orthoptera: Acrididae). *Heredity*, **35**:375–387. doi:10.1038/hdy.1975.108.

Hewitt GM (1988). Hybrid zones-natural laboratories for evolutionary studies. *Trends in Ecology & Evolution*, **3**:158–67. doi:10.1016/0169-5347(88)90033-X.

Hewitt GM (2000). The genetic legacy of the Quaternary ice ages. *Nature*, **405**:907–13. doi:10.1038/35016000.

Hewitt GM (2001). Speciation, hybrid zones and phylogeography - or seeing genes in space and time. *Molecular Ecology*, **10**:537–549. doi:10.1046/j.1365-294x.2001.01202.x.

Hewitt GM (2004). Genetic consequences of climatic oscillations in the Quaternary. *Philosophical Transactions of the Royal Society of London Series B, Biological sciences*, **359**:183–95. doi:10.1098/rstb.2003.1388.

Hewitt GM (2011). Quaternary phylogeography: the roots of hybrid zones. *Genetica*, **139**:617–638. doi:10.1007/s10709-011-9547-3.

Hewitt GM & John B (1972). Inter-population sex chromosome polymorphism in the grasshopper *Podisma pedestris*. *Chromosoma*, **37**:23–42. doi:10.1007/BF00329555.

Hewitt GM, Mason P, & Nichols RA (1989). Sperm precedence and homogamy across a hybrid zone in the alpine grasshopper *Podisma pedestris*. *Heredity*, **62**:343–353. doi:10.1038/hdy.1989.49.

Hewitt GM, Nichols RA, & Barton NH (1987). Homogamy in a Hybrid Zone in the Alpine Grasshopper *Podisma pedestris*. *Heredity*, **59**:457–466. doi:10.1038/hdy.1987.156.

Huxley J (1938). Clines: an Auxiliary Taxonomic Principle. *Nature*, **142**:219–220. doi:10.1038/142219a0.

Jiang H, Lei R, Ding SW, & Zhu S (2014). Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics*, **15**:182. doi:10.1186/1471-2105-15-182.

John B & Hewitt GM (1970). Inter-population sex chromosome polymorphism in the grasshopper *Podisma pedestris*. *Chromosoma*, **31**:291–308. doi:10.1007/BF00321226.

Keightley PD, Ness RW, Halligan DL, & Haddrill PR (2014). Estimation of the spontaneous mutation rate per nucleotide site in a *Drosophila melanogaster* full-sib family. *Genetics*, **196**:313–20. doi:10.1534/genetics.113.158758.

Keightley PD, Pinharanda A, Ness RW, Simpson F, Dasmahapatra KK, Mallet J, Davey JW, & Jiggins CD (2015). Estimation of the Spontaneous Mutation Rate in *Heliconius melpomene*. *Molecular Biology and Evolution*, **32**:239–243. doi:10.1093/molbev/msu302.

Keller I, Veltsos P, & Nichols RA (2008). The frequency of rDNA variants within individuals provides evidence of population history and gene flow across a grasshopper hybrid zone. *Evolution*, **62**:833–844. doi:10.1111/j.1558-5646.2008.00320.x.

Kelly JK & Noor MAF (1996). Speciation by Reinforcement: A Model Derived from Studies of *Drosophila*. *Genetics*, **143**:1485–1497.

Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC, Kielbasa SM, Wan R, Sato K, Horton P, & Frith MC (2011). Adaptive seeds tame genomic sequence comparison. *Genome Research*, **21**:487–493. doi:10.1101/gr.113985.110.

Kirkpatrick M & Guerrero RF (2014). Signatures of Sex-Antagonistic Selection on Recombining Sex Chromosomes. *Genetics*, **197**:531–541. doi:10.1534/genetics.113.156026.

Klopfstein S, Currat M, & Excoffier L (2006). The Fate of Mutations Surfing on the Wave of a Range Expansion. *Molecular Biology and Evolution*, **23**:482–490. doi:10.1093/molbev/msj057.

Kowalczyk M, Tatsuta H, Grzywacz B, & Warchałowska-Śliwa E (2009). Relationship Between Chromosomal Races/Subraces in the Brachypterous Grasshopper *Podisma sapporensis* (Orthoptera: Acrididae) Inferred from Mitochondrial ND2 and COI Gene Sequences. *Annals of the Entomological Society of America*, **101**:837–844.

Langmead B & Salzberg SL (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**:357–359. doi:10.1038/nmeth.1923.

Lansman RA, Shade RO, Shapira JF, & Avise JC (1981). The use of restriction endonucleases to measure mitochondrial DNA sequence relatedness in natural populations. *Journal of Molecular Evolution*, **17**:214–226.

Lasne C, Sgrò CM, & Connallon T (2017). The Relative Contributions of the X Chromosome and Autosomes to Local Adaptation. *Genetics*, **205**:1285–1304. doi:10.1534/genetics.116.194670.

Li L, Stoeckert CJ, & Roos DS (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research*, **13**:2178–89. doi:10.1101/gr.1224503.

Liou LW & Price TD (1994). Speciation by reinforcement of premating isolation. *Evolution*, **48**:1451–1459. doi:10.1111/j.1558-5646.1994.tb02187.x.

Liu S, Wei W, Chu Y, Zhang L, Shen J, & An C (2014). De Novo Transcriptome Analysis of Wing Development-Related Signaling Pathways in *Locusta migratoria* Manilensis and *Ostrinia furnacalis* (Guenée). *PLoS ONE*, **9**:e106 770. doi:10.1371/journal.pone.0106770.

Lohse K, Chmelik M, Martin SH, & Barton NH (2016). Efficient Strategies for Calculating Blockwise Likelihoods Under the Coalescent. *Genetics*, **202**:775–786. doi:10.1534/genetics.115.183814.

Lohse K, Harrison RJ, & Barton NH (2011). A General Method for Calculating Likelihoods Under the Coalescent Process. *Genetics*, **189**:977–987. doi:10.1534/genetics.111.129569.

Lopez JV, Yuhki N, Masuda R, Modi W, & O'Brien SJ (1994). Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. *Journal of Molecular Evolution*, **39**:174–190. doi:10.1007/BF00163806.

Mao X, Dong J, Hua P, He G, Zhang S, & Rossiter SJ (2014). Heteroplasmy and Ancient Translocation of Mitochondrial DNA to the Nucleus in the Chinese Horseshoe Bat (*Rhinolophus sinicus*) Complex. *PLoS ONE*, **9**:e98 035. doi:10.1371/journal.pone.0098035.

Martínez-Rodríguez P, Hernández-Pérez M, & Bella JL (2013). Detection of Spiroplasma and Wolbachia in the Bacterial Gonad Community of *Chorthippus parallelus*. *Microbial Ecology*, **66**:211–223. doi:10.1007/s00248-013-0226-z.

Mason PL, Nichols RA, & Hewitt GM (1995). Philopatry in the alpine grasshopper, *Podisma pedestris*: a novel experimental and analytical method. *Ecological Entomology*, **20**:137–145. doi:10.1111/j.1365-2311.1995.tb00439.x.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, & DePristo MA (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, **20**:1297–303. doi:10.1101/gr.107524.110.

Meisel RP & Connallon T (2013). The faster-X effect: integrating theory and data. *Trends in Genetics*, **29**:537–544. doi:10.1016/J.TIG.2013.05.009.

Mohandesan E, Speller CF, Peters J, Uerpmann HP, Uerpmann M, De Cupere B, Hofreiter M, & Burger PA (2017). Combined hybridization capture and shotgun sequencing for ancient DNA analysis of extinct wild and domestic dromedary camel. *Molecular Ecology Resources*, **17**:300–313. doi:10.1111/1755-0998.12551.

Nei M (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences of the United States of America*, **70**:3321–3323. doi:10.1073/pnas.70.12.3321.

Nichols RA & Hewitt GM (1986). Population structure and the shape of a chromosomal cline between two races of *Podisma pedestris* (Orthoptera: Acrididae). *Biological Journal of the Linnean Society*, **29**:301–316. doi:10.1111/j.1095-8312.1986.tb00282.x.

Nichols RA & Hewitt GM (1988). Genetical and ecological differentiation across a hybrid zone. *Ecological Entomology*, **13**:39–49. doi:10.1111/j.1365-2311.1988.tb00331.x.

Nichols RA & Hewitt GM (1994). The genetic consequences of long distance dispersal during colonization. *Heredity*, **72**:312–317. doi:10.1038/hdy.1994.41.

Novák P, Neumann P, Pech J, Steinhaisl J, & Macas J (2013). RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, **29**:792–793. doi:10.1093/bioinformatics/btt054.

Nürnberger B, Lohse K, Fijarczyk A, Szymura JM, & Blaxter ML (2016). Para-allopatry in hybridizing fire-bellied toads (*Bombina bombina* and *B. variegata*): Inference from transcriptome-wide coalescence analyses. *Evolution*, **70**:1803–1818. doi:10.1111/evo.12978.

Paradis E (2010). pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics*, **26**:419–420. doi:10.1093/bioinformatics/btp696.

Rice WR (1987). The Accumulation of Sexually Antagonistic Genes as a Selective Agent Promoting the Evolution of Reduced Recombination between Primitive Sex Chromosomes. *Evolution*, **41**:911–914. doi:10.2307/2408899.

Rice WR (1996). Sexually antagonistic male adaptation triggered by experimental arrest of female evolution. *Nature*, **381**:232–4. doi:10.1038/381232a0.

Rice WR (1998). Male fitness increases when females are eliminated from gene pool: implications for the Y chromosome. *Proceedings of the National Academy of Sciences of the United States of America*, **95**:6217–21.

Robinson MR, Pilkington JG, Clutton-Brock TH, Pemberton JM, & Kruuk LEB (2006). Live fast, die young: trade-offs between fitness components and sexually antagonistic selection on weaponry in soay sheep. *Evolution*, **60**:2168. doi:10.1554/06-128.1.

Straub SCK, Parks M, Weitemier K, Fishbein M, Cronn RC, & Liston A (2012). Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. *American Journal of Botany*, **99**:349–64. doi:10.3732/ajb.1100335.

Suyama M, Torrents D, & Bork P (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic acids research*, **34**:W609–12. doi:10.1093/nar/gkl315.

Swenson NG & Howard DJ (2005). Clustering of contact zones, hybrid zones, and phylogeographic breaks in North America. *The American naturalist*, **166**:581–91. doi:10.1086/491688.

Szymura JM & Barton NH (1991). The genetic structure of the hybrid zone between the fire-bellied toads *Bombina bombina* AND *B. variegata*: Comparisons between transects and between loci. *Evolution*, **45**:237–261. doi:10.1111/J.1558-5646.1991.TB04400.X.

Takahata N, Satta Y, & Klein J (1995). Divergence Time and Population Size in the Lineage Leading to Modern Humans. *Theoretical Population Biology*, **48**:198–221. doi:10.1006/TPBI.1995.1026.

Tatsuta H, Hoshizaki S, Bugrov AG, Warchalowska-Sliwa E, Tatsuki S, & Akimoto S (2006). Origin of Chromosomal Rearrangement: Phylogenetic Relationship Between X0/XX and XY/XX Chromosomal Races in the Brachypterous Grasshopper *Podisma sapporensis* (Orthoptera: Acrididae). *Annals of the Entomological Society of America*, **99**:457–462.

Tu X, Wang J, Hao K, Whitman DW, Fan Y, Cao G, & Zhang Z (2015). Transcriptomic and proteomic analysis of pre-diapause and non-diapause eggs of migratory locust, *Locusta migratoria* L. (Orthoptera: Acridoidea). *Scientific Reports*, **5**:11 402. doi:10.1038/srep11402.

Twyford AD & Ness RW (2016). Strategies for complete plastid genome sequencing. *Molecular Ecology Resources*. doi:10.1111/1755-0998.12626.

van der Valk T, Lona Durazo F, Dalén L, & Guschanski K (2017). Whole mitochondrial genome capture from faecal samples and museum-preserved specimens. *Molecular Ecology Resources*. doi:10.1111/1755-0998.12699.

Varkonyi PL & Domokos G (2006). Static Equilibria of Rigid Bodies: Dice, Pebbles, and the Poincare-Hopf Theorem. *Journal of Nonlinear Science*, **16**:255–281. doi:10.1007/s00332-005-0691-8.

Veltsos P, Keller I, & Nichols RA (2008). The inexorable spread of a newly arisen neo-Y chromosome. *PLoS genetics*, **4**:e1000 082. doi:10.1371/journal.pgen.1000082.

Veltsos P, Keller I, & Nichols RA (2009). Geographically localised bursts of ribosomal DNA mobility in the grasshopper *Podisma pedestris*. *Heredity*, **103**:54–61. doi:10.1038/hdy.2009.32.

Virdee SR & Hewitt GM (1994). Clines for Hybrid Dysfunction in a Grasshopper Hybrid Zone. *Evolution*, **48**:392. doi:10.2307/2410100.

Warchałowska-Śliwa E, Bugrov AG, Sugano Y, Maryańska-Nadachowska A, & Akimoto SI (2008). Experimental hybridization between X0 and XY chromosome races in the grasshopper *Podisma sapporensis* (Orthoptera: Acrididae). II. Cytological analysis of embryos and adults of F1 and F2 generations. *European Journal of Entomology*, **105**:45–52.

Weir BS & Cockerham CC (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, **38**:1358–1370. doi:10.2307/2408641.

Westerman M, Barton NH, & Hewitt GM (1987). Differences in DNA content between two chromosomal races of the grasshopper *Podisma pedestris*. *Heredity*, **58**:221–228. doi:10.1038/hdy.1987.36.

Zhang DX & Hewitt GM (1996). Nuclear integrations: challenges for mitochondrial DNA markers. *Trends in Ecology & Evolution*, **11**:247–251. doi:10.1016/0169-5347(96)10031-8.

Appendix A

This appendix corresponds to Chapter 2.

Data retrieval and formatting

An ASCII file with elevation data was downloaded and cropped to the area of interest using the julia script `elev.jl`:

```
# open gzipped ASCII file of elevations
# retrieved from http://srtm.csi.cgiar.org/SELECTION/listImages.asp
# tile covering 44.5-44.0N, 5.8-7.0E
using GZip
f = GZip.open("srtm\_38\_04.asc.gz", "r")
# the file has 6 header lines which will be ignored
# there are 6001 lines and 6001 columns
# make corresponding array
elevs = Array(Int64, 6001, 6001)
# parse data file by line
i = 0
for line in eachline(f)
i += 1
if i > 6
println(i)
elevs[i - 6,:] = [parse(Int64, i) for i in split(line)]
end
end
# utility functions translating lat/long into array indices
function lati(x::Float64)
return(Int(round((45. - x) * 6001 / 5.)))
end
function loti(x::Float64)
return(Int(round((x - 5.) * 6001 / 5.)))
end
# of all the data imported save only the area of interest:
interest = elevs[lati(44.5):lati(44.0), loti(5.8):loti(7.0)]
writecsv("interest_test.csv", interest)
# end of program
```

Inference of watersheds

Mathematica (version 11.2) was then used to infer the position of watersheds (elev_appendix.nb):

```
(*Inferring watersheds from elevation data
Constants and utility functions*)
elevs =Import["/home/hannes/Dropbox/elevations/interest.csv","CSV"];
Dimensions[elevs]
{423,964}
(*Convert elevations to image. ImageAdjust rescales values so
they are between 0 and 1*)
img=Image[elevs]//ImageAdjust
(*Determine position of watersheds using starting points
What localities do the starting points correspond to?*)
sheds=WatershedComponents[Image[img],{{270,423-400},{682,423-400},
{20,423-400},{900,423-400}}];
Image[sheds]
Export["sheds.csv",a]
```

Appendix B

This appendix corresponds to Chapter 3.

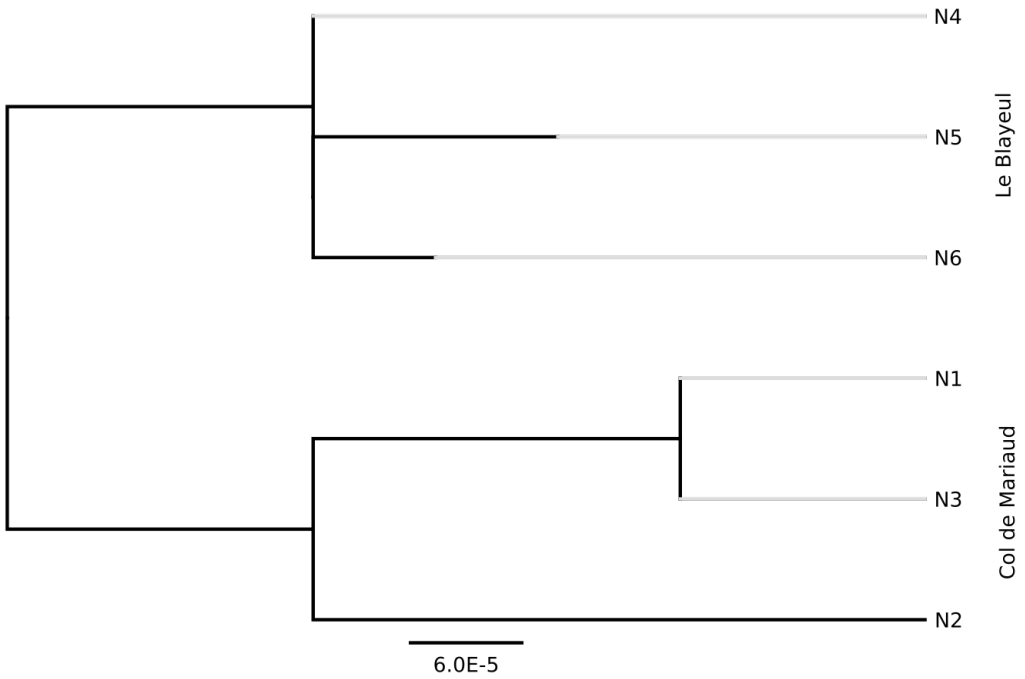

Figure 27: **Unrooted neighbour-joining tree of six mitochondrial genome sequences** generated from Illumina NextSeq short-read data.

Appendix C

This appendix corresponds to Chapter 5.

Table 9: **Sample information for the sequence capture experiment.** The columns are: Sample number, number of individuals in the sample, sex, karyotype, Sampling location, populations label, name of the GPS point, Latitude, and Longitude.

Smpl	Inds	Sex	Kar	Location	Pop	GPS	Lat	Long
1	1	U	M	Crête des Mélèzes	15#2	122	44.36065	6.23195
2	1	U	M	Crête des Mélèzes	15#2	122	44.36065	6.23195
3	1	U	M	Crête des Mélèzes	15#2	122	44.36065	6.23195
4	1	U	M	Crête des Mélèzes	15#2	122	44.36065	6.23195
5	1	U	M	Crête des Mélèzes	15#2	122	44.36065	6.23195
6	1	U	M	Dormillouse	16#13	176	44.40842	6.38546
7	1	U	M	Dormillouse	16#13	176	44.40842	6.38546
8	1	U	M	Dormillouse	16#13	176	44.40842	6.38546
9	1	U	M	Dormillouse	16#13	176	44.40842	6.38546
10	1	U	M	Dormillouse	16#13	176	44.40842	6.38546
11	1	U	F	Dormillouse	16#13	176	44.40842	6.38546
12	1	U	F	Dormillouse	16#13	176	44.40842	6.38546
13	1	U	F	Dormillouse	16#13	176	44.40842	6.38546
14	1	U	F	Dormillouse	16#13	176	44.40842	6.38546
15	1	U	F	Dormillouse	16#13	176	44.40842	6.38546
16	1	F	M	Col de Mariaud	16#1	80	44.27766	6.42942
17	1	F	M	Col de Mariaud	16#1	80	44.27766	6.42942
18	1	F	M	Col de Mariaud	16#1	80	44.27766	6.42942
19	1	F	M	Col de Mariaud	16#1	80	44.27766	6.42942
20	1	F	M	Col de Mariaud	16#1	80	44.27766	6.42942
21	1	F	F	Col de Mariaud	16#1	80	44.27766	6.42942
22	1	F	F	Col de Mariaud	16#1	80	44.27766	6.42942
23	1	F	F	Col de Mariaud	16#1	80	44.27766	6.42942
24	1	F	F	Le Blayeul	16#14	77	44.26657	6.30616
25	1	F	F	Le Blayeul	16#14	77	44.26657	6.30616
26	1	F	F	Le Blayeul	16#14	77	44.26657	6.30616
27	1	F	F	Le Blayeul	16#14	77	44.26657	6.30616
28	1	F	F	Le Blayeul	16#14	77	44.26657	6.30616
29	1	F	M	Le Blayeul	16#14	77	44.26657	6.30616
30	1	F	M	Le Blayeul	16#14	77	44.26657	6.30616
31	1	F	M	Le Blayeul	16#14	77	44.26657	6.30616

Smpl	Inds	Sex	Kar	Location	Pop	GPS	Lat	Long
32	1	F	M	Le Blayeul	16#14	77	44.26657	6.30616
33	1	F	M	Le Blayeul	16#14	77	44.26657	6.30616
34	5	F	M	La Bigue	15#42-45	77	44.26657	6.30616
35	3	F	F	La Bigue	15#43-45	77	44.26657	6.30616
					15#42	160	44.14161	6.19186
					15#43	161	44.14212	6.19002
					15#44	162	44.14637	6.18747
					15#45	163	44.12802	6.19852
36	1	F	M	Lac d'Allos	16#19	180	44.22544	6.70695
37	1	F	M	Lac d'Allos	16#19	180	44.22544	6.70695
38	1	F	M	Lac d'Allos	16#19	180	44.22544	6.70695
39	1	F	M	Lac d'Allos	16#19	180	44.22544	6.70695
40	1	F	M	Lac d'Allos	16#19	180	44.22544	6.70695
41	1	F	F	Lac d'Allos	16#19	180	44.22544	6.70695
42	1	F	F	Lac d'Allos	16#19	180	44.22544	6.70695
43	1	F	F	Lac d'Allos	16#19	180	44.22544	6.70695
44	1	F	F	Lac d'Allos	16#19	180	44.22544	6.70695
45	1	F	F	Lac d'Allos	16#19	180	44.22544	6.70695
46	1	U	M	Lac d'Allos	16#21	182	44.24837	6.71654
47	1	U	M	Lac d'Allos	16#21	182	44.24837	6.71654
48	1	U	M	Lac d'Allos	16#21	182	44.24837	6.71654
49	1	U	M	Lac d'Allos	16#21	182	44.24837	6.71654
50	1	U	M	Lac d'Allos	16#21	182	44.24837	6.71654
51	1	U	F	Lac d'Allos	16#21	182	44.24837	6.71654
52	1	U	F	Lac d'Allos	16#21	182	44.24837	6.71654
53	1	U	F	Lac d'Allos	16#21	182	44.24837	6.71654
54	1	U	F	Lac d'Allos	16#21	182	44.24837	6.71654
55	1	U	F	Lac d'Allos	16#21	182	44.24837	6.71654
56	5	T	M	l'Aiguillette	16#6	169	44.32445	6.44556
57	5	T	M	l'Aiguillette	16#7	170	44.32801	6.44687
58	5	T	M	l'Aiguillette	16#15	>200m from 177		
59	5	T	M	l'Aiguillette	16#16	177	44.34712	6.445
60	5	T	F	l'Aiguillette	16#6	169	44.32445	6.44556
61	5	T	F	l'Aiguillette	16#7	170	44.32801	6.44687
62	5	T	F	l'Aiguillette	16#15	>200m from 177		
63	5	T	F	l'Aiguillette	16#16	177	44.34712	6.445
64	5	TU	F	W of Les Tomples	16#3	165	44.33321	6.26421
65	5	T	M	Lac d'Allos	16#18	179	44.23033	6.70405
66	2	T	M	Lac d'Allos	16#20	181	44.23261	6.71275

Smpl	Inds	Sex	Kar	Location	Pop	GPS	Lat	Long
67	3	T	M	Lac d'Allos	16#22	183	44.24132	6.70751
68	5	T	F	Lac d'Allos	16#18	179	44.23033	6.70405
69	4	T	F	Lac d'Allos	16#22	183	44.24132	6.70751
70	5	U	M	Col Bas	16#12	175	44.38638	6.40138
71	5	U	F	Col Bas	16#12	175	44.38638	6.40138
72	5	F	M	Roger's Cabin	16#5	167	44.31817	6.27965
73	4	F	F	Roger's Cabin	16#5	167	44.31817	6.27965
74	3	F	F	Roger's Cabin	16#5	167	44.31817	6.27965
75	3	T	F	E of Les Tomples	16#4	166	44.321	6.27217
76	3	T	F	E of Les Tomples	16#4	166	44.321	6.27217
77	5	T	F	C Les Tomples	16#23	153	44.32809	6.26962
78	5	T	F	C Les Tomples	16#23	153	44.32809	6.26962
79	5	T	F	C Les Tomples	16#24	155	44.32733	6.26886
80	5	T	F	C Les Tomples	16#24	155	44.32733	6.26886
81	5	T	F	C Les Tomples	16#25	184	44.32731	6.27013
82	5	T	F	C Les Tomples	16#25	184	44.32731	6.27013
83	5	T	M	Barles	15#14	135	44.27524	6.21057
84	5	T	M	Barles	15#15	136	44.27444	6.2101
85	5	T	M	Barles	15#16	137	44.27347	6.20969
86	5	T	M	Barles	15#17	138	44.27253	6.20964
87	5	T	M	Barles	15#18	139	44.27168	6.20936
88	5	T	M	Barles	15#19	140	44.27062	6.20967
89	5	T	M	Barles	15#24	144	44.2698	6.20835
90	5	T	M	Barles	15#26	147	44.26891	6.20895
91	5	T	M	Barles	14#T12	93	44.2679	6.2099
92	5	T	M	Barles	14#T11	92	44.26709	6.20991
93	5	T	M	Barles	14#T10	91	44.26627	6.21068
94	5	T	M	Barles	14#T9	90	44.26609	6.21164
95	5	T	M	Barles	14#T85	89	44.26583	6.21209
96	5	T	M	Barles	14#T5	95	44.26394	6.21557
97	5	T	M	Barles	14#T4	96	44.26446	6.21659
98	5	T	M	Barles	14#T3	97	44.26467	6.21779
99	5	T	M	Barles	14#T2	98	44.26462	6.21901
100	5	T	M	Barles	15#22/28	143	44.26396	6.22137
101	5	T	M	Barles	15#29	148	44.26385	6.22264
102	5	T	M	Barles	15#30	149	44.26365	6.22388
103	5	T	M	Barles	15#31	150	44.26337	6.22507
104	5	T	M	Barles	15#32	151	44.26338	6.22633
105	5	T	M	Barles	15#33	152	44.26341	6.22698

Smpl	Inds	Sex	Kar	Location	Pop	GPS	Lat	Long
106	1	T	M	Barles	14#T7	87	44.26506	6.21299
107	1	T	M	Barles	14#T7	87	44.26506	6.21299
108	1	T	M	Barles	14#T7	87	44.26506	6.21299
109	1	T	M	Barles	14#T7	87	44.26506	6.21299
110	1	T	M	Barles	14#T7	87	44.26506	6.21299
111	1	T	M	Les Tomples	14#U2	100	44.3282	6.26324
112	1	T	M	Les Tomples	14#U2	100	44.3282	6.26324
113	1	T	M	Les Tomples	14#U2	100	44.3282	6.26324
114	1	T	M	Les Tomples	14#U2	100	44.3282	6.26324
115	1	T	M	Les Tomples	14#U2	100	44.3282	6.26324
116	1	T	M	Les Tomples	14#U3	102	44.32815	6.26308
117	1	T	M	Les Tomples	14#U3	102	44.32815	6.26308
118	1	T	M	Les Tomples	14#U3	102	44.32815	6.26308
119	1	T	M	Les Tomples	14#U3	102	44.32815	6.26308
120	1	T	M	Les Tomples	14#U3	102	44.32815	6.26308
121	1	T	M	Les Tomples	14#U4	103	44.32782	6.26396
122	1	T	M	Les Tomples	14#U4	103	44.32782	6.26396
123	1	T	M	Les Tomples	14#U4	103	44.32782	6.26396
124	1	T	M	Les Tomples	14#U4	103	44.32782	6.26396
125	1	T	M	Les Tomples	14#U4	103	44.32782	6.26396
126	1	T	M	Les Tomples	14#U5	104	44.32755	6.2649
127	1	T	M	Les Tomples	14#U5	104	44.32755	6.2649
128	1	T	M	Les Tomples	14#U5	104	44.32755	6.2649
129	1	T	M	Les Tomples	14#U5	104	44.32755	6.2649
130	1	T	M	Les Tomples	14#U5	104	44.32755	6.2649
131	1	T	M	Les Tomples	14#U6	105	44.32727	6.26608
132	1	T	M	Les Tomples	14#U6	105	44.32727	6.26608
133	1	T	M	Les Tomples	14#U6	105	44.32727	6.26608
134	1	T	M	Les Tomples	14#U6	105	44.32727	6.26608
135	1	T	M	Les Tomples	14#U6	105	44.32727	6.26608
136	1	T	M	Les Tomples	14#U7	106	44.32701	6.26702
137	1	T	M	Les Tomples	14#U7	106	44.32701	6.26702
138	1	T	M	Les Tomples	14#U7	106	44.32701	6.26702
139	1	T	M	Les Tomples	14#U7	106	44.32701	6.26702
140	1	T	M	Les Tomples	14#U7	106	44.32701	6.26702
141	1	T	M	Les Tomples	14#U8	107	44.3266	6.26826
142	1	T	M	Les Tomples	14#U8	107	44.3266	6.26826
143	1	T	M	Les Tomples	14#U8	107	44.3266	6.26826
144	1	T	M	Les Tomples	14#U8	107	44.3266	6.26826

Smpl	Inds	Sex	Kar	Location	Pop	GPS	Lat	Long
145	1	T	M	Les Tomples	14#U8	107	44.3266	6.26826
146	1	T	M	Les Tomples	14#U9	108	44.32622	6.2694
147	1	T	M	Les Tomples	14#U9	108	44.32622	6.2694
148	1	T	M	Les Tomples	14#U9	108	44.32622	6.2694
149	1	T	M	Les Tomples	14#U9	108	44.32622	6.2694
150	1	T	M	Les Tomples	14#U9	108	44.32622	6.2694
151	1	T	M	Les Tomples	14#U10	109	44.32592	6.27052
152	1	T	M	Les Tomples	14#U10	109	44.32592	6.27052
153	1	T	M	Les Tomples	14#U10	109	44.32592	6.27052
154	1	T	M	Les Tomples	14#U10	109	44.32592	6.27052
155	1	T	M	Les Tomples	14#U10	109	44.32592	6.27052
156	1	T	M	Les Tomples	14#U11	110	44.32556	6.27165
157	1	T	M	Les Tomples	14#U11	110	44.32556	6.27165
158	1	T	M	Les Tomples	14#U11	110	44.32556	6.27165
159	1	T	M	Les Tomples	14#U11	110	44.32556	6.27165
160	1	T	M	Les Tomples	14#U11	110	44.32556	6.27165
161	1	T	M	Les Tomples	14#U12	112	44.32525	6.27281
162	1	T	M	Les Tomples	14#U12	112	44.32525	6.27281
163	1	T	M	Les Tomples	14#U12	112	44.32525	6.27281
164	1	T	M	Les Tomples	14#U12	112	44.32525	6.27281
165	1	T	M	Les Tomples	14#U12	112	44.32525	6.27281
166	1	T	M	Les Tomples	16#4	166	44.321	6.27217
167	1	T	M	Les Tomples	16#4	166	44.321	6.27217
168	1	T	M	Les Tomples	16#4	166	44.321	6.27217
169	1	T	M	Les Tomples	16#4	166	44.321	6.27217
170	1	T	M	Les Tomples	16#4	166	44.321	6.27217
171	1	T	M	W of Les Tomples	16#3	165	44.33321	6.26421
172	1	T	M	W of Les Tomples	16#3	165	44.33321	6.26421
173	1	T	M	W of Les Tomples	16#3	165	44.33321	6.26421
174	1	T	M	W of Les Tomples	16#3	165	44.33321	6.26421
175	1	T	M	W of Les Tomples	16#3	165	44.33321	6.26421
176	5		M	Authon	15#10	131	44.22238	6.17136
177	5		M	Authon	15#6	127	44.24883	6.1713
178	1	F	F	Col de Mariaud	cross	80	44.27766	6.42942
179	1	U	M	Col Bas	cross	175	44.38638	6.40138
180	1	na	na	cross	cross			
181	1	na	na	cross	cross			
182	1	na	na	cross	cross			
183	1	T	M	Barles	14#T8	88	44.26555	6.21255

Smpl	Inds	Sex	Kar	Location	Pop	GPS	Lat	Long
184	1	T	M	Barles	14#T8	88	44.26555	6.21255
185	1	T	M	Barles	14#T8	88	44.26555	6.21255
186	1	T	M	Barles	14#T8	88	44.26555	6.21255
187	1	T	M	Barles	14#T8	88	44.26555	6.21255
188	1	T	M	Barles	14#T6	86	44.26448	6.21457
189	1	T	M	Barles	14#T6	86	44.26448	6.21457
190	1	T	M	Barles	14#T6	86	44.26448	6.21457
191	1	T	M	Barles	14#T6	86	44.26448	6.21457
192	1	T	M	Barles	14#T6	86	44.26448	6.21457

Appendix D – CV – Hannes Becher

Academic history

2018–2021	Postdoc with Alex Twyford, UoE
2018	Postdoc with Brian Charlesworth, UoE
2014–2017	PhD with Richard Nichols, QMUL
2016–2017	Visiting Research Student with Konrad Lohse, UoE
2013–2014	Research Assistant with Andrew Leitch, QMUL
2010–2012	MSc, Syst Bot and Cytogenetics, MLU Halle-Wittenberg
2007–2010	BSc, Biology (Palynology), MLU Halle-Wittenberg

Skills/Experience

Drylab: HTS assembly, mapping and variant calling; statistical and mathematical modelling; programming in R, python, julia, and Mathematica

Wetlab: DNA-extraction, PCR, library preparation for HTS, cytogenetics (FisH), microscopy

Teaching: demonstrating in undergraduate courses, support on final year and MSc projects, replacement lecturing for cytogenetics

Fieldwork: annual expeditions to the Alpine *Podisma pedestris* hybrid zone

Grants/Awards

- Genetics Society Training Grant £1000
- Genetics Society Grant for the organisation of EMPSEB21 £2500
- Leonardo-da-Vinci-Fellowship €6000

Publications

1. Winterfeld G, **Becher H**, Voshell S & Hilu K (2018). Karyotype evolution in *Phalaris* (Poaceae): The role of reductional dysploidy, polyploidy and

chromosome alteration in a wide-spread and diverse genus. *PLOS ONE*, 13, e0192869.

2. Wang W, Ma L, **Becher H**, Garcia S, Kovaříkova A, Leitch IJ, Leitch AR & Kovařík A (2016). Astonishing 35S rDNA diversity in the gymnosperm species *Cycas revoluta* Thunb. *Chromosoma*, 125, 683–699.

3. Ma L, Hatlen A, Kelly LJ, **Becher H**, Wang W, Kovařík A, Leitch IJ, & Leitch AR (2015). Angiosperms are unique amongst land plant lineages in the occurrence of key genes in the RNA dependent DNA methylation (RdDM) pathway. *Genome Biology and Evolution*, 7, 2648–2662.

4. Winterfeld G, Schneider J, **Becher H**, Dickie J, & Röser M (2015). Karyosystematics of the Australasian stipoid grass *Austrostipa* and related genera: chromosome sizes, ploidy, chromosome base numbers, and phylogeny. *Australian Systematic Botany*, 28, 145–159.

5. **Becher H**, Ma L, Kelly LJ, Kovařík A, Leitch IJ, & Leitch AR (2014). Endogenous pararetrovirus sequences associated with 24 nt small RNAs at the centromeres of *Fritillaria imperialis* L. (Liliaceae), a species with a giant genome. *The Plant Journal*, 80, 823–833.

6. Hellmund M, Wennrich V, **Becher H**, Krichel A, Bruelheide H, & Melles M (2011). Zur Vegetationsgeschichte im Umkreis des Süßen Sees, Lkr. Mansfeld-Südharz – Ergebnisse von Pollen- und Elementanalysen. In: Bork HJ, Meller H, Gerlach R (eds), *Umweltarchäologie – Naturkatastrophen und Umweltwandel im archäologischen Befund, Tagungen des Landesmuseums für Vorgeschichte Halle (Saale)* 6, Halle, 111–127.

Appendix E

This paper was published in the first weeks of my PhD. While working on this project I learned scripting in R and python and how to use RepeatExplorer, which was later useful for the analyses carried out for Chapter 3.

the plant journal

SEB Society for Experimental Biology

The Plant Journal (2014) **80,** 823–833 doi: 10.1111/tpj.12673

Endogenous pararetrovirus sequences associated with 24 nt small RNAs at the centromeres of *Fritillaria imperialis* L. (Liliaceae), a species with a giant genome

Hannes Becher[1], Lu Ma[1], Laura J. Kelly[1,2], Ales Kovarik[3], Ilia J. Leitch[2] and Andrew R. Leitch[1,*]

[1]*School of Biological and Chemical Sciences, Queen Mary University of London, London E1 4NS, UK,*

[2]*Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3AB, UK, and*

[3]*Institute of Biophysics, Academy of Sciences of the Czech Republic, Brno CZ–61265, Czech Republic*

Received 20 June 2014; revised 10 September 2014; accepted 11 September 2014; published online 17 September 2014.

*For correspondence (e-mail a.r.leitch@qmul.ac.uk).

SUMMARY

Endogenous pararetroviral sequences are the most commonly found virus sequences integrated into angiosperm genomes. We describe an endogenous pararetrovirus (EPRV) repeat in *Fritillaria imperialis*, a species that is under study as a result of its exceptionally large genome (1C = 42 096 Mbp, approximately 240 times bigger than *Arabidopsis thaliana*). The repeat (FriEPRV) was identified from Illumina reads using the RepeatExplorer pipeline, and exists in a complex genomic organization at the centromere of most, or all, chromosomes. The repeat was reconstructed into three consensus sequences that formed three interconnected loops, one of which carries sequence motifs expected of an EPRV (including the *gag* and *pol* domains). FriEPRV shows sequence similarity to members of the *Caulimoviridae* pararetrovirus family, with phylogenetic analysis indicating a close relationship to *Petuvirus*. It is possible that no complete EPRV sequence exists, although our data suggest an abundance that exceeds the genome size of Arabidopsis. Analysis of single nucleotide polymorphisms revealed elevated levels of C→T and G→A transitions, consistent with deamination of methylated cytosine. Bisulphite sequencing revealed high levels of methylation at CG and CHG motifs (up to 100%), and 15–20% methylation, on average, at CHH motifs. FriEPRV's centromeric location may suggest targeted insertion, perhaps associated with meiotic drive. We observed an abundance of 24 nt small RNAs that specifically target FriEPRV, potentially providing a signature of RNA-dependent DNA methylation. Such signatures of epigenetic regulation suggest that the huge genome of *F. imperialis* has not arisen as a consequence of a catastrophic breakdown in the regulation of repeat amplification.

Keywords: pararetrovirus, *Fritillaria imperialis*, centromere, RNA interference, cytosine methylation, RdDM, giant genome.

INTRODUCTION

Viral sequences from two virus families, the *Caulimoviridae* (pararetroviruses) and the *Geminiviridae*, are known to have integrated into the nuclear genomes of angiosperms. These endogenous viral sequences are typically present in a few hundred or thousand copies, contributing to the repetitive genome content. They show varying levels of degradation and rearrangements, and, in most cases, are functionally defective. Nevertheless, there are a few examples of endogenous pararetroviral sequences (EPRVs) belonging to *Caulimoviridae* for which complete copies exist, or for which recombination may generate complete copies. Under certain circumstances (e.g. stress, hybridization), these EPRVs may become circularized and released from the genome and become infective (Teycheney and Geering, 2011; Chabannes and Iskra-Caruana, 2013).

Repetitive DNA amplification processes and the efficiency of repeat removal are generally accepted to account for genome size differences in angiosperms of the same ploidy, for example in cotton (*Gossypium hirsutum*; Hawkins *et al.*, 2009), *Arabidopsis lyrata* (Hu *et al.*, 2011) and *Nicotiana tabacum* (Renny-Byfield *et al.*, 2011). In recent years, it has become increasingly clear that much of the repetitive DNA in the genome is epigenetically regulated via pathways that are mediated by small interfering RNAs (siRNAs), leading to DNA methylation, histone modification and chromatin remodelling (Simon and Meyers, 2011; Kim and Zilberman, 2014; Matzke and Mosher, 2014). Our

understanding of these pathways is primarily derived from studies on species with small genome sizes (e.g. *Arabidopsis thaliana*, 1C = 157 Mbp) (Bennett *et al.*, 2003), in which epigenetic regulation silences transposable elements and thus hampers their proliferation. However, angiosperms have an exceptionally large (2400-fold) range in genome size (Leitch and Leitch, 2012, 2013), and the question arises as to whether epigenetic pathways also operate in species with such large genomes, as failure of epigenetic regulation may potentially lead to genome enlargement (Kelly and Leitch, 2011).

With the advent of high-throughput next-generation sequencing approaches, species with larger genomes have recently become amenable to study, and are now the focus of research into the origin and evolution of genome obesity (Kelly and Leitch, 2011; Kelly *et al.*, 2012; Nystedt *et al.*, 2013). Even using low-coverage sequencing data, it is possible to reliably analyse the repetitive content of large genomes as shown for pea (*Pisum sativum*) by Macas *et al.* (2007). Here we use low-coverage genomic DNA sequence data from Illumina's next-generation sequencing platform to characterize an EPRV in *Fritillaria imperialis* (FriEPRV), a species with an exceptionally large genome (1C = 42 096 Mbp or more, if B chromosomes are present) (Leitch *et al.*, 2007). In addition, we also characterize the small RNA (smRNA) and cytosine methylation status of the *F. imperialis* genome, to obtain insights into the regulation of FriEPRV and its evolution.

RESULTS

Characterization of a pararetrovirus-like sequence in *F. imperialis*

We used RepeatExplorer to characterize the repetitive DNA content of the genome of *F. imperialis* (see Experimental procedures, Novák *et al.*, 2010, 2013). The EPRVs identified, hereafter called FriEPRV, were in the 15th largest repeat cluster in terms of the number of reads it contains. The cluster comprises 7727 reads (0.386%) of the 2 000 842 analysed, the latter representing a genome coverage of 0.475% in *F. imperialis*. The 7727 reads were assembled into 109 contigs, 14 of which comprised more than 30 reads (Table 1). These 14 contigs are AT-rich and in total contain 7119 reads, corresponding to 92.1% of all reads in the cluster. All contigs of FriEPRV are contained in Supporting Data file S1. The cluster graph of FriEPRV consists of three distinct loops (loops 1–3, Figure 1a), each of which comprises a set of overlapping or similar sequence reads. The region at which the loops connect occurs because there are similar sequences in the three loops. In addition, multiple Illumina read pairs (insert size 300–500 bp) span the loops, providing evidence that the loops are physically connected (Figure 1b).

Table 1 Characterization of contigs and reads obtained in FriEPRV: contigs in bold were not included in the loop consensus sequences

Consensus sequence (length in bp)	Contig name	Number of reads	Contig length in bp	Read depth	GC content in %
Loop 1 (5939)	8	90	250	36.00	42.1
	18	2035	2072	98.21	34.9
	29	1110	1174	94.55	37.6
	39	269	424	63.44	34.0
	80	1690	1948	86.76	35.7
	90	**63**	**253**	**24.90**	**34.4**
	92	79	210	37.62	38.1
	94	629	877	71.72	40.6
Loop 2 (1322)	7	44	286	15.38	43.4
	38	311	969	32.09	41.7
	78	195	647	30.14	41.4
Loop 3 (1216)	11	203	543	37.38	46.2
	102	**34**	**248**	**13.71**	**44.4**
	105	367	932	39.38	42.6

The 14 contigs reconstructed in FriEPRV (Table 1) were further assembled using Geneious (http://geneious.com/) to generate consensus sequences for loops 1–3 (Figure 2 and Table 1). Read depth analysis across each consensus sequence revealed a read depth of approximately 70–130-fold for loop 1, approximately 20–40-fold for loop 2 and approximately 40–60-fold for loop 3 (Figure 2). Thus loop 1 of FriEPRV has approximately twice the read depth of the other loops. Dot-plot analyses of the consensus sequences revealed a minisatellite at the 5′ and 3′ ends of each consensus (Figure S1), probably causing RepeatExplorer to link reads into three loops (Figure 1). Before downstream analysis, this satellite was removed from all consensus sequences except the 5′ end of loop 1. Use of Tandem Repeats Finder (Benson, 1999) showed that the minisatellite has a 30 bp consensus motif (AAGGGGG TTTTGATGCTCTAATACCACTCG)$_n$.

For the consensus sequences of loops 2 and 3, BLASTn and BLASTx analysis revealed no significant matches against National Center for Biotechnology Information databases (using an *e*–value threshold of $<10^{-5}$). In contrast, BLASTx analysis of the consensus sequence of loop 1 revealed a match against a predicted polyprotein domain in *Cicer arietinum*. There were also matches against a *Citrus* endogenous pararetrovirus, a hypothetical protein in *Eutrema salsugineum*, petunia vein clearing virus (PVCV) and a predicted protein from *Populus trichocarpa*. BLASTx also detected conserved domains characteristic of retrovirus-like sequences. All domains identified are illustrated in Figure 2. The consensus sequence of loop 1 also contains a large open reading frame (ORF) of 5361 bp, which is terminated at the 3′ end by a double stop codon (TAATAA). FriEPRV's gen-

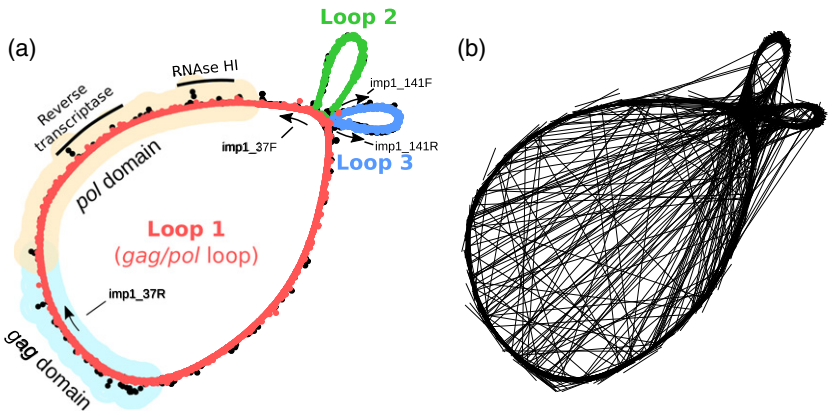

Figure 1. Graphical 2D projection of the structure of FriEPRV generated using RepeatExplorer (Novák *et al.*, 2010).
(a) Each node (dot) represents one of the 7727 Illumina sequence reads. The placement of the nodes reflects sequence similarity, with sequences that are most alike being placed closest together. The sequences have been shaded according to their presence in one of three loops: large loop 1 (red) and minor loops 2 and 3 (blue and green). Black dots represent reads that are not incorporated in the contigs mentioned in Table 1. The protein-coding domains were annotated in RepeatExplorer with reference to RepeatMasker and Repbase (Jurka *et al.*, 2005) and the conserved domain database (Marchler-Bauer *et al.*, 2011), and by MEGABLAST and BLASTx analysis. The labelled arrows indicate the position of primers used for PCR (see Table S1 for details).
(b) Illumina read pairs are connected by lines, indicating physical connection of the loops.

ome proportion of 0.386% equates to 162 Mbp (see Appendix S1 for details of the confidence interval of this estimate). As repeat clustering of 0.475% of the genome of *F. imperialis* results in an approximately 100-fold coverage for loop 1 (*gag*/*pol* loop, 5939 bp), we assume the presence of approximately 21 000 copy equivalents of loop 1 of FriEPRV. Appendix S2 provides a more detailed explanation of how the copy number estimates were derived.

Previously, Llorens *et al.* (2009) analysed the phylogenetic relationships between members of the pararetrovirus family *Caulimoviridae*. Given the sequence similarities that we found, we aligned the consensus sequence of FriEPRV loop 1 to representative *Caulimoviridae* sequences from Llorens *et al.* (2009). Phylogenetic analysis of amino acid sequences from *Caulimoviridae* revealed a strongly supported relationship between FriEPRV and PVCV (Figure S2). Dot-plot analysis of amino acid sequences of FriEPRV loop 1 against PVCV also revealed high levels of conservation in regions with sequence similarity to the *gag* and *pol* polyprotein genes, which also occur in LTR retrotransposons and related elements, as in other members of the *Caulimoviridae* (Figure 3). *pol* usually encodes protease, reverse transcriptase and ribonuclease. In transposable elements, it also encodes integrase (Llorens *et al.*, 2009). Further sequence similarities were found between the N–terminal regions of PVCV and FriEPRV (movement protein, Figure 3), although the DNA sequences diverge between the *gag*/*pol* and N–terminal regions (corresponding to amino acids 300–800 in Figure 3), even when considering potential frame shifts.

Small RNAs

The smRNA reads were mapped to the consensus sequences of the three FriEPRV loops (Figure 2, black bars). Of 4609 mapped smRNA reads, the majority (67%) were found to belong to the 24 nt size category (Figure S3). Peaks of high smRNA read abundance (>100 reads) were found at particular positions along the three consensus sequences. Six peaks were observed on loop 1, one of which was in the *pol* domain. One of the highest peaks corresponds to the 30 bp minisatellite (Figure 2, asterisk). None of the smRNA reads that mapped to FriEPRV mapped to any other repeat cluster identified by RepeatExplorer in the full *F. imperialis* dataset of 2 000 842 reads.

Cytosine methylation

After mapping Illumina reads from sodium bisulfite-treated genomic DNA to each of the three consensus sequences of FriEPRV, we calculated the proportion of cytosines at each location that were methylated in the genomic DNA (see Experimental procedures). We observed that, in the context of CG and CHG (where H represents A, T or C), most cytosine residues were methylated, with medians of approximately 86 and 90% of cytosines at any particular position (Figure 4a). In contrast, cytosines in the context of CHH were significantly less methylated, with methylation being found at only approximately 11% of cytosines at any particular position in the consensus sequences. Nevertheless, some cytosines in the CHH context were methylated in most mapped reads (up to 88%) (see Figure 4a). We compared the distribution of methylated cytosines in the context of CHH (1641 sites) against their map position in

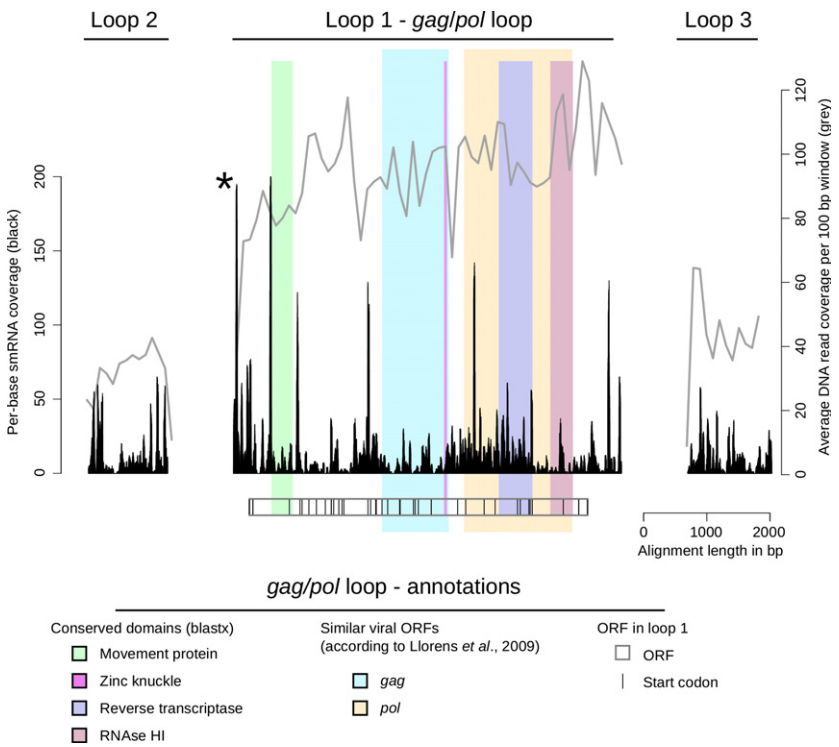

Figure 2. Read depth analysis (per base pair) using Illumina reads of genomic DNA (grey line) and small RNAs (black bars) against the consensus sequences for loops 1–3.
The box under loop 1 indicates the largest open reading frame (ORF); vertical bars represent in-frame start codons. Protein domains were annotated as described by Llorens *et al.* (2009) and BLASTx results. Note that the read depth from genomic DNA for loops 2 and 3 is approximately half that observed in loop 1. Note the multiple peaks of small RNA reads, one of which corresponds to a 30 bp mini-satellite (indicated by an asterisk).

FriEPRV sequences (210 bp windows). Forty cytosines in the context of CHH were predominantly methylated (>70%), these cytosines were scattered apparently at random across loops 1–3. Loop 1 showed comparatively little methylation in the CHH context at its 5′ end (Figure 4b).

An analysis of single nucleotide polymorphisms (SNPs) revealed that C→T and G→A transitions were most abundant (Figure S4), a pattern consistent with deamination of methylated cytosines.

Fluorescence *in situ* and Southern hybridization

PCR primers were designed to amplify the region containing the *pol* domain of loop 1 and the majority of loop 3 (Figure 1a and Table S1). Gel electrophoresis revealed a range of product sizes in each case, probably reflecting heterogeneity amongst natural FriEPRV sequences (shown for loop 1 in Figure 5a). Thus the loop 1 and 3 probes used for fluorescent *in situ* hybridization (FISH) each contained a range of size fragments.

The karyotype of *F. imperialis* comprises 2 $n = 24$ chromosomes, 20 of which are acrocentric, two of which are sub-metacentric and two of which are metacentric (see asterisks in Figure 6a). The *F. imperialis* individual depicted in Figure 6 also contained one additional B chromosome in each metaphase (see arrow in Figure 6a). The Texas Red-labelled loop 1 probe (carrying the *gag*/*pol* domain) hybridized strongly to the centromeric region of the 20 acrocentric chromosomes and the two sub-metacentric chromosomes. In addition, weak labelling was sometimes observed at the centromere of one metacentric chromosome (but is not visible in Figure 6a), although its homologue always appeared unlabelled, even when the image was contrast-enhanced (result not shown). The B chromosome showed a very small and weak hybridization signal in the telomeric region.

The Texas Red-labelled loop 1 probe (red signal) and the Alexa Fluor 488-labelled loop 3 probe (green signal) co-localized at metaphase, giving rise to a yellow hybridization signal (results not shown), a pattern that was also observed for the majority of signals at interphase (Figure 6b). At interphase, some FriEPRV sequences were more decondensed than others, and, some had stronger loop 1 than loop 3 signals, resulting in red rather than yellow florescence (see arrows in Figure 6b).

Southern hybridization, using the probe for loop 1 on *Bst*NI-restricted genomic DNA from a range of *Fritillaria* species (selected to reflect the phylogenetic diversity of the genus) showed only a weak hybridization signal in the lane for *F. imperialis* (Figure 5b). The long exposure times required to reveal the signal reflect the relatively low genome proportion (approximately 0.4%) of FriEPRV. There was one prominent hybridization band at approximately 3 kb, and a number of indistinct minor bands, potentially reflecting multiple copies of similar sequence. None of the other species (listed in Table S2) showed any signal even though the same membrane probed for 18S rDNA revealed signal in all lanes (Figure 5c).

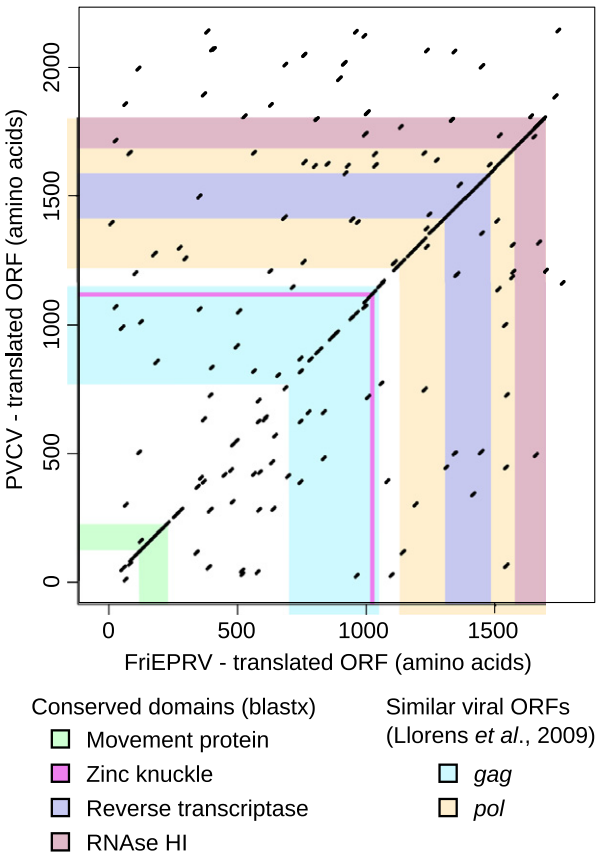

Figure 3. Dot-plot analysis of the translated consensus sequence of FriEPRV loop 1 and petunia vein clearing virus (PVCV).
Note the similarity in the protein-coding domains, particularly in the *pol* domain. Domains are annotated as shown in Figure 2. Note the less conserved region between amino acids 300 and 800.

DISCUSSION

FriEPRV: an endogenous pararetroviral sequence

Our data are indicative of the relatively recent insertion and amplification of an endogenous pararetroviral sequence in *Fritillaria* that we have called FriEPRV. Evidence for this includes the findings that (i) a Southern hybridization signal was detected only in *F. imperialis*, (ii) there is conservation of coding sequences in a dot-plot analysis with PVCV, and (iii) there is an absence of stop codons in the ORF of FriEPRV. The consensus sequence of FriEPRV loop 1 reveals domains with similarity to the *gag* and *pol* domains of retroviruses and LTR retrotransposons (Figure 2). BLAST analysis, sequence reconstruction and phylogenetic analysis all indicate that FriEPRV loop 1 is an EPRV sequence, which, in our analysis (Figure S2), is most closely related to PVCV, which belongs to the genus *Petuvirus* within the *Caulimoviridae*. Like PVCV, FriEPRV exhibits one large polyprotein ORF containing *gag* and *pol* domains (Noreen *et al.*, 2007). Dot-plot analysis comparing the ORF of FriEPRV loop 1 with that of PVCV (Figure 3) revealed considerable sequence similarity at the amino acid level, with the exception of the region between amino acids 300 and 800. In LTR retrotransposons, which are closely related to pararetroviruses, this region contains an integrase domain that is not functional in pararetroviruses (Chabannes *et al.*, 2013).

Plant pararetroviruses are the most abundant viral sequences integrated into plant genomes, and have previously been reported in species belonging to Asteraceae, Asparagaceae, Bromeliaceae, Musaceae, Poaceae, Rosaceae, Rutaceae, Salicaceae, Solanaceae and Vitaceae

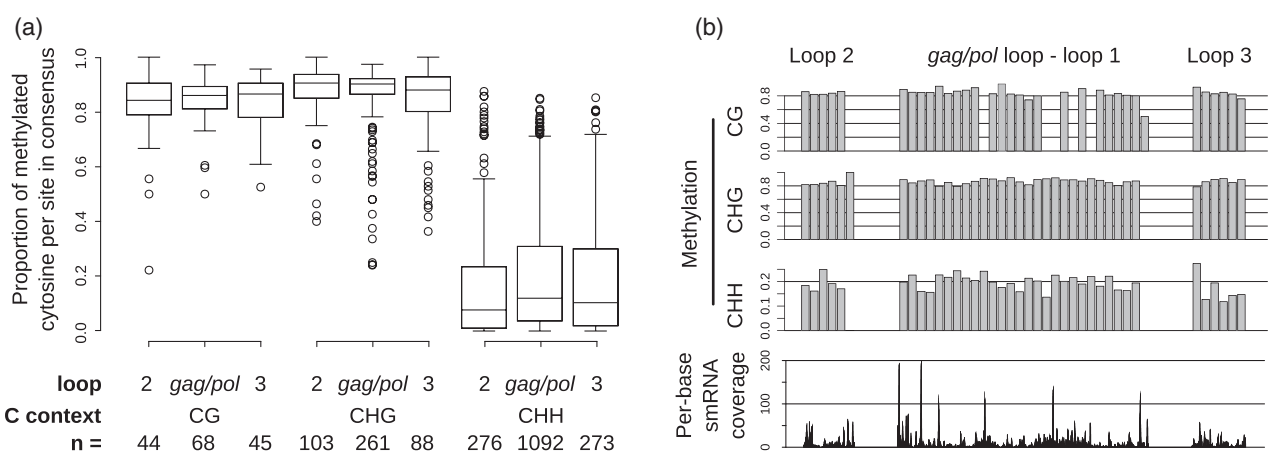

Figure 4. Analysis of the frequency and distribution of methylated cytosines in the consensus sequences of loops 1–3 of FriEPRV.
(a) The numbers (*n*) of CG, CHG and CHH motifs in the consensus sequences of loops 1–3 are shown at the bottom. For each cytosine in the consensus, the proportion of reads that are modified (and hence unmethylated) or unaltered (and hence methylated) by sodium bisulfite treatment was recorded and is indicated as the methylation proportion at each cytosine. Only cytosines with more than ninefold bisulfite read coverage were considered. The plot shows the range of methylation ratios in the context of CG, CHG and CHH motifs. For each box plot, the lower, middle and upper horizontal lines represent the first, second and third quartiles, respectively. The maximum whisker length is 1.5 times the interquartile range. Results for cytosines outside this range are indicated as circles.
(b) The mean level of methylation along the sequences of loops 1–3 is shown in windows of 210 bp for all cytosine contexts (CG, CHG and CHH). In loop 1, there are some regions without cytosine in CG context, hence the missing bars. Per-base coverage of smRNA mapped to the loop consensus sequences is shown at the bottom for comparison.

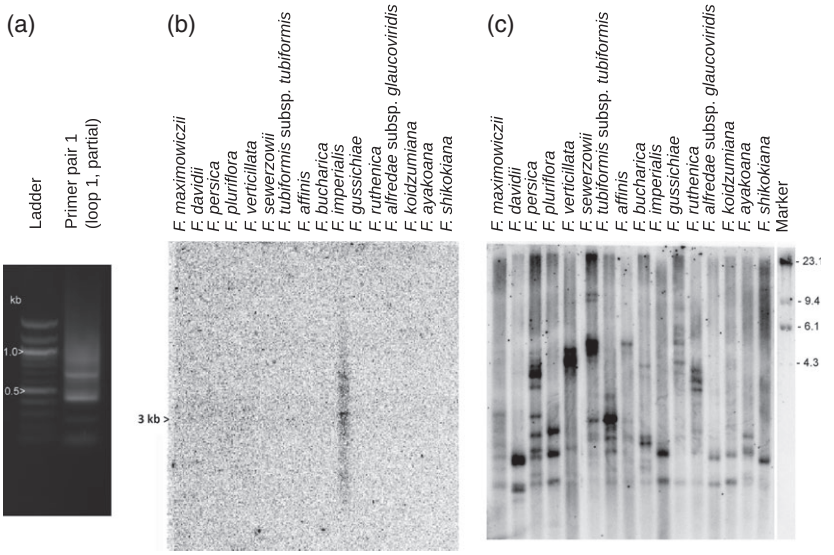

Figure 5. Occurrence of FriEPRV.
(a) Gel electrophoresis analysis of PCR products using primer pair 1 (see Table S1) against loop 1 of FriEPRV for *Fritillaria imperialis*. Note the smear indicating a range of sizes of products.
(b) The products in (a) were used as a probe for Southern hybridization of *Bst*NI-digested genomic DNA from a range of *Fritillaria* species. Only *F. imperialis* shows three hybridization bands in addition to a smear.
(c) Re-hybridization of the same membrane using the 18S probe. In each species, the probe hybridized to bands of variable size, reflecting length polymorphisms in the intergenic spacer.

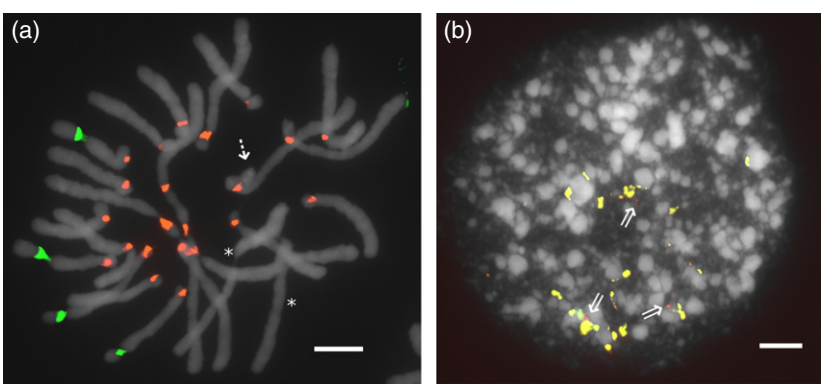

Figure 6. FISH analysis in root tip cells of *F. imperialis*. DNA was stained with 4′,6–diamidino-2–phenylindole (white) to detect the EPRV sequence FriEPRV.
(a) Metaphase chromosomes probed with 18S rDNA (green fluorescence) and loop 1 probe (red fluorescence). Note the centromeric location of the FriEPRV sequences, as indicated by the occurrence of a constriction in many of the signals, on acrocentric and sub-metacentric chromosomes. There are small signals on the B chromosome (arrow), and an absence of signal on the two metacentric chromosomes (asterisks).
(b) Interphase nucleus probed with loop 3 probe (green fluorescence) and loop 1 probe (red). Note that, in most cases, the signals are co-localized (giving a yellow fluorescence signal), although there are a few signals that carry only loop 1 label (arrows). Note also the variable condensation state of FriEPRV at interphase. Scale bars = 10 μm.

(Teycheney and Geering, 2011; Chabannes and Iskra-Caruana, 2013). We are unaware of other reports of an endogenous pararetrovirus in a species belonging to Liliaceae, or indeed in the whole of the monocot order Liliales comprising approximately 1300 species.

Organization of the pararetrovirus sequence FriEPRV

Functional EPRVs may occur as complete elements (which occur in tandem arrays in *Petunia hybrida*) or as incomplete, fragmented and rearranged sequences either at a single locus, as in *Musa balbisiana*, or at several genomic loci, as in *Nicotiana* × *edwardsonii* (Chabannes and Iskra-Caruana, 2013). Despite such differences in genome organization, inter-specific hybridization, stress or wounding may result in release of full-length circularized infectious viral genomes, potentially involving complex patterns of recombination (Chabannes and Iskra-Caruana, 2013). In the present study, RepeatExplorer reconstructed FriEPRV as a complete viral genome (loop 1). Although our analysis utilizes short reads and covers only a fraction of the *F. imperialis* genome, there is evidence indicating the genome organization of FriEPRV: (i) loop 1 is approximately twice as long as the two other loops, (ii) all three loops share similar terminal sequences, and (iii) the loops are physically connected as evidenced by the presence of Illumina read pairs with mates in two loops (see Figure 1b). These results agree with a fundamental organization of the type $(\text{loop 1, 2, 1, 3})_n$. Potentially loops 2 and 3 provide spacer sequences between FriEPRV copies, or contain promoter domains for their transcription. At present, we do not know

whether a complete copy of FriEPRV exists in the genome. Nevertheless, the large range of fragment sizes generated by PCR using primer pair 1 to amplify loop 1 FriEPRV sequences (Figure 5a), and the range of restriction fragment sizes that hybridized with the loop 1 probe following Southern hybridization (Figure 5b), suggest either a complex organization of FriEPRV in the *F. imperialis* genome, including truncated variants, and/or much sequence divergence between the repeats. Nevertheless, given that it takes only one complete EPRV element, or recombination of fragmented elements, to generate active viral genomes (Chabannes and Iskra-Caruana, 2013; Chabannes *et al.*, 2013), it is possible that the FriEPRV in *F. imperialis* may have the potential to be infectious, even if all individual copies are rearranged.

We predict that the equivalent of approximately 21 000 copies of FriEPRV are present in the genome of *F. imperialis* (see Appendix S2). While abundant ERPVs have been reported for *Nicotiana tabacum* (Jakowitsch *et al.*, 1999), *Solanum lycopersicum* (Staginnus *et al.*, 2007) and *Capsicum annuum* (Kim *et al.*, 2014), *F. imperialis* exhibits the highest copy number reported for an ERPV to date. Together, FriEPRV sequences are estimated to account for approximately 162 Mbp of the *F. imperialis* genome, which is comparable to the genome size of *Arabidopsis thaliana* (Bennett *et al.*, 2003). However in the context of *F. imperialis*, this copy number represents a genome proportion of only approximately 0.4%.

Chromosomal distribution of FriEPRV

Fluorescent *in situ* hybridization (FISH) revealed that sequences corresponding to loops 1 and 3 of FriEPRV were predominantly located at the morphological centromeres of the A chromosomes. Additional but very weak signals were also seen on B chromosomes (Figure 6a). Although EPRV sequences at centromeric/pericentromeric sites have been reported previously in *Solanum lycopersicum* (tomato), *Solanum habrochaites* (Staginnus *et al.*, 2007) and *Solanum tuberosum* (potato, Hansen *et al.*, 2005), they are not restricted to such regions in these species. The only other example in which EPRV sequences were found to be restricted to the centromeres was for PVCV in *Petunia hybrida* (Richert-Pöggeler *et al.*, 2003), and it is perhaps notable that PVCV is phylogenetically the closest known EPRV sequence to FriEPRV (Figure S2).

What explains FriEPRV's localization at the centromeres? It is unlikely there is selection against widespread integration of the sequence, except perhaps in genic domains, given the enormous genome size of *F. imperialis* (1C = 42 096 Mbp). Previously, Richert-Pöggeler and Shepherd (1997) annotated the region between the movement protein and the RNA-binding domain in PVCV as integrase. This region has diverged between FriEPRV and PVCV (Figure 3), and other pararetroviruses neither possess functional integrases nor do they require integration for replication (Chabannes *et al.*, 2013). However, the centromeric location of both PVCV and FriEPRV reveals accumulation of the repeats at the centromeres, perhaps even targeted integration, as reported for other transposable elements (Neumann *et al.*, 2011).

Malik and Henikoff (2009) suggested a model for centromeric repeat proliferation involving meiotic drive: the non-Mendelian segregation of chromosomes bearing certain sequences during meiosis leading to their preferential inheritance (Kanizay and Dawe, 2009; Kanizay *et al.*, 2013). If the presence of FriEPRV sequences at centromeres causes (or at some point during its evolution has caused) meiotic drive, it is likely that FriEPRV will spread (or have spread), as chromosomes with more centromeric FriEPRV sequences will be preferentially inherited. This has been reported for heterochromatin in maize (*Zea mays*) and centromeric repeats in other grasses (Kanizay and Dawe, 2009; Kanizay *et al.*, 2013). It also agrees with the theory of 'boom–burst' cycles in the evolution of centromere sequences (Zhang *et al.*, 2014).

Epigenetic silencing in the context of a giant genome

Our knowledge about epigenetic silencing and its implications for genome composition is mainly based on the analysis of species with small genomes. For example, in Arabidopsis, epigenetic silencing of repetitive DNA is regulated by two pathways: (i) the self-reinforcing maintenance methylation pathway involving the KRYPTONITE (KYP) family of histone methyl transferases and chromomethylases, and (ii) the sRNA-directed DNA methylation pathway. KRYPTONITE/chromomethylase maintenance methylation pathways are chiefly active in the centres of long transposable elements (TEs) with compact chromatin structures, in which KRYPTONITE proteins dimethylate lysine 9 of histone 3 (H3K9me2) in the presence of DNA methylation at CHG and CHH motifs. Chromomethylase proteins in turn recognize H3K9me2 sites and methylate DNA, keeping the central parts of TEs silenced independent of their actual sequence. In contrast, the sequences at the edges of TEs are mainly silenced by sRNA-directed DNA methylation, which is sequence-specific and does not depend on H3K9me2 and chromatin compaction. Here, sequence specificity matters, because sequences neighbouring silenced TEs may need to be transcriptionally active and hence retain a more open euchromatin conformation (Kim and Zilberman, 2014).

For TEs surrounded by euchromatic areas in Arabidopsis, the sRNA-directed DNA methylation pathway often causes characteristic peaks of CHH methylation and smRNA abundance at the edges of TEs, while CHG methylation tends to be lower at the edges of TEs and to increase and then plateau towards the centres (Zemach *et al.*, 2013). Our results for FriEPRV show that it is methylated almost

equally throughout its sequence (Figure 4b), while smRNA mapping shows several peaks across both coding and non-coding domains (Figures 2 and 4b). This pattern resembles the epigenetic landscape observed for *Gypsy* elements in Arabidopsis, which are mainly located in heterochromatic areas (Zemach *et al.*, 2013). However, we were unable to find any association between smRNA peaks and patterns of cytosine methylation.

The presence of SNPs (Figure S4) suggests that the FriEPRV sequences have not evolved in a concerted manner (as some arrays of repetitive sequences do) but have diverged. This is consistent with its comparatively high level of cytosine methylation in the CHG context, which is probably connected to condensed chromatin and hence a reduced frequency of the recombination required for homogenization processes (Peng and Karpen, 2008). Recombination-based processes are not only the basis for concerted evolution (Nei and Rooney, 2005) but also removal of repetitive sequences and hence genome downsizing (Grover and Wendel, 2010). The fact that most of the SNPs are C→T transitions probably reflects the deamination of methylated cytosine, which is indicative of long-term methylation of FriEPRV since its amplification in this species.

Overall, our analysis of the structure, methylation status and epigenetic regulation of FriEPRV clearly indicates that, at least for this genomic sequence in *F. imperialis*, epigenetic regulation is operational. These data thus provide direct evidence that the huge genomes of *F. imperialis* are not caused by a catastrophic breakdown in the epigenetic regulation of repeats leading to their runaway amplification.

EXPERIMENTAL PROCEDURES

Plant material and root preparations

The plant species studied are listed in Table S2. All species were grown at the Royal Botanic Gardens, Kew, UK, in pots on open ground, or were obtained from the personal collection of Laurence Hill (www.fritillariaicones.com). Root tips were collected and placed in a saturated solution of α–bromonaphtalene for 24 h on ice to accumulate cells at metaphase. The roots were then fixed in 3:1 v/v ethanol/glacial acetic acid, and incubated for 3 h at room temperature (approximately 20°C). Finally, roots were transferred to 100% ethanol and stored at −20°C.

DNA/RNA extraction

Total genomic DNA was extracted from silica-dried leaf tissue using a cetyl trimethylammonium bromide method as described by Kovarík *et al.* (2000). DNA samples were resuspended in 1× TE buffer, treated with RNase A and purified using NucleoSpin® Extract II columns (Macherey-Nagel, http://www.mn-net.com/).

Total RNA, including the small RNA fraction (smRNA, <200 nt), was extracted from mature leaf tissue that had been stored in RNAlater® (Qiagen, http://www.qiagen.com/) at −80°C. Extractions were performed using a mirVana™ miRNA isolation kit (Life Technologies, https://www.lifetechnologies.com) according to the manufacturer's instructions.

Illumina sequencing

Paired-end sequencing (2× 100 bp, 300–500 bp insert size) of total genomic DNA, sodium bisulfite-treated genomic DNA (unmethylated cytosine converted to uracil), and the smRNA fraction was performed by the Centre for Genomic Research at the University of Liverpool on the Illumina (http://www.illumina.com/) HiSeq® platform. smRNA libraries were size-selected to retain fragments with inserts of 6–66 bp. Sequencing data were supplied in FASTQ format with adaptors already trimmed.

Bioinformatics

RepeatExplorer. The RepeatExplorer pipeline (http://repeatexplorer.umbr.cas.cz/) clusters next-generation sequencing reads into groups of similar reads, and assembles contigs from these reads. The output may also be presented in graphical form by application of a Fruchterman–Reingold algorithm, which positions reads that are most similar closest together (Novák *et al.*, 2010). RepeatExplorer also annotates cluster regions using RepeatMasker and Repbase (Jurka *et al.*, 2005) and the conserved domain database (Marchler-Bauer *et al.*, 2011).

After removing reads with a Phred score of less than 20 for more than 10% of their bases, 2 000 842 Illumina paired-end reads of genomic DNA from *F. imperialis* (corresponding to 0.475% of the genome) were clustered into repeat families using RepeatExplorer (Novák *et al.*, 2010, 2013), as described previously (Renny-Byfield *et al.*, 2013).

Consensus sequences. To generate consensus sequences for the three looped domains of cluster FriEPRV (see Results), contigs containing ≥30 reads generated by RepeatExplorer (see Table 1) were further assembled manually through similarities in overlapping domains using Geneious® version 5.5.6 (http://www.geneious.com/). Contigs had low coverage at their ends. These low-coverage regions were removed if they showed ambiguities compared to high-coverage regions. Contigs 90 and 102, which contained divergent sequences at low coverage, were removed completely. The remainder were assembled into three consensus sequences corresponding to the three domains of FriEPRV: loop 1, loop 2 and loop 3. To further characterize these consensus sequences, MEGABLAST and BLASTx searches were performed using the National Center for Biotechnology Information online BLAST facility (http://blast.ncbi.nlm.nih.gov/Blast.cgi) with an *e*-value threshold of 10^{-5}.

Annotation of SNPs. We used the CLC Genomics Workbench 6.5.1 (CLCbio, http://www.clcbio.com/) to map all 2 000 842 genomic reads against the loop consensus sequences, and subsequently detected SNPs using CLC's function 'Probabilistic Variant Detection' with default settings.

Methylation analysis. The distribution of cytosine methylation in FriEPRV was analysed through comparison of sequences from sodium bisulfite-treated genomic DNA with consensus sequences from untreated DNA using Bismark (Krueger and Andrews, 2011), which comprises a set of Perl scripts that align bisulfite-modified FASTQ reads from Illumina data to reference sequences using the Bowtie short read aligner (Langmead *et al.*, 2009). Of 39 287 390 reads, approximately 0.1% could be mapped to the loop consensus sequences. For each cytosine in the consensus sequence, we recorded the number of bisulfite-treated reads with a cytosine or thymine at each cytosine position in the consensus, with thymine

arising as a consequence of bisulfite treatment of unmethylated cytosine. In order to avoid artefacts arising from low coverage, we considered only cytosine residues that were more than ninefold covered with reads derived from bisulfite-treated DNA. We excluded cytosine residues involved in SNPs (see Appendix S3). For each remaining cytosine in the consensus, we calculated the percentage of times that it was methylated in genomic DNA. These percentages indicate how many individual copies of FriEPRV are methylated at any particular cytosine.

Phylogenetic analysis. A 5361 nt open reading frame within FriEPRV loop 1 (see Results for details) was identified, translated and added to sequences from the GAGCOATPOL_caulimoviridae dataset (http://gydb.org/index.php/Collection_alignments), which contains genic domain sequences of the *Caulimoviridae* family. The amino acid sequences were aligned using T–Coffee (Notredame *et al.*, 2000), using the program's default parameters at the T–Coffee server (www.tcoffee.org/) (Di Tommaso *et al.*, 2011). Phylogenetic reconstruction using maximum parsimony and bootstrap analysis were performed using PAUP* version 4.0 b10 (Swofford, 2003), as described by Kelly *et al.* (2013); only positions scored as 'good' in the T–Coffee alignment were included in the phylogenetic analyses. Details of alignment scoring are given at www.tcoffee.org/Documentation/t_coffee/t_coffee_tutorial.htm.

RNA mapping. smRNA reads were quality-filtered using Biopython (http://biopython.org/, only Phred scores >30 allowed), and converted to FASTA format. smRNA reads were mapped to the three loop consensus sequences of FriEPRV using Bowtie (Langmead *et al.*, 2009), allowing only perfect matches over the whole read sequence.

PCR

The primer pairs, designed using PerlPrimer version .1.21 with default settings (downloaded from http://perlprimer.sourceforge.net/download.html), and the PCR conditions are shown in Table S1, and their positions are shown in Figure 1. PCR was performed in 20 μl volumes containing 1 unit of NEB® Taq polymerase (NEB, https://www.neb.com/), 1× NEB® Taq buffer, 200 μM of each nucleotide and approximately 20 ng *F. imperialis* DNA, with the addition of 0.8 μl formamide for 18S PCR.

Fluorescent *in situ* hybridization

Cell spreads. Cell spreads were prepared from fixed root tips as described by Lim *et al.* (2006). Briefly, root tips were lightly digested using 1% v/v pectinase and 2% v/v cellulase in citrate buffer, and spread under a coverslip in 60% v/v acetic acid. Coverslips were then removed after freezing in liquid nitrogen.

Probe labelling. PCR products (see above) were checked using agarose gel electrophoresis, and then purified and eluted in water using a QIAquick PCR purification kit (Qiagen). Purified PCR products were labelled by nick translation in 40 μl volumes containing 3 μg DNA, 0.01 M β–mercaptoethanol, 50 μM of each dATP, dGTP and dCTP, 10 μM dTTP, 50 μM labelled dUTP (Texas Red for the loop 1 probe; Alexa Fluor 488 for the18S and loop 3 probes), 50 mM Tris/HCl, 5 mM $MgCl_2$, 0.005% w/v BSA, 20 units DNA polymerase I (Invitrogen, http://www.lifetechnologies.com/) and 0.2 units DNase I (Thermo Scientific, http://www.thermoscientific.com/). Nick translation was performed at 15°C for 1 h. Products were checked via agarose gel electrophoresis.

FISH. Fluorescent *in situ* hybridization (FISH) was performed as described by Lim *et al.* (2006) with minor modifications. Briefly, slides were rinsed twice for 5 min in 2× SSC (0.3 M sodium chloride and 0.03 M tri-sodium citrate), fixed for 10 min in 4% v/v formaldehyde in 2× SSC, rinsed four times for 3 min in 2× SSC, dehydrated in ethanol of increasing concentrations (70, 95 and 100%) for 2 min each, and finally air-dried. Then 20 μl of hybridization mix containing 50% w/v formamide, 0.3 M sodium chloride, 0.03 M tri-sodium citrate, 10 mM Tris/HCl, 2 mM EDTA, 2.8 μg salmon sperm DNA (for blocking) and 75–150 ng probe were added to each slide. Slides were denatured at 80°C using a Dyad™ DNA engine carrying a PRINS block (Bio–Rad, http://www.bio–rad.com/). After incubation in an airtight plastic box at 37°C overnight, slides were quickly rinsed in 2× SSC at room temperature to remove coverslips, incubated in 2× SSC at 55°C for 20 min, briefly rinsed in 2× SSC again, dehydrated in ethanol as described above, and air-dried. Slides were mounted in a drop of Vector Shield™ mounting medium containing 4′,6–diamidino-2–phenylindole (Vector Laboratories, https://www.vectorlabs.com/), and stored for at least 1 h before microscope analysis.

Images were taken using a DMRA2 epifluorescence microscope (Leica, http://www.leica-microsystems.com/) equipped with an Orca ER™ monochrome camera (Hamamatsu, http://www.hamamatsu.com/). Contrast and brightness enhancement as well as merging of layers were performed using OpenLab™ imaging software (Improvision, Cambridge, UK, http://www.perkinelmer.co.uk/pages/020/cellularimaging/improvision/default.xhtml).

Southern hybridization. Southern hybridization was performed as described previously (Koukalova *et al.*, 2010). Extractions of total genomic DNA were sourced from the DNA bank at the Royal Botanic Gardens, Kew, UK (see Table S2); species of *Fritillaria* were selected to represent the phylogenetic diversity of the genus (Day *et al.*, 2014). DNA was digested with excess *Bst*NI (2× 2 h), and subjected to electrophoresis in agarose gels, using 1–2 μg of DNA/lane. Gels were blotted onto Hybond N+ membrane (GE Healthcare Life Sciences, http://www.gelifesciences.com/). After transfer, Southern blot hybridization was performed in 0.25 M sodium phosphate buffer, pH 7.0, supplemented with 7% w/v SDS at 65°C for 16 h, using α–^{32}P-dCTP-labelled PCR products from primer pair 1 (>10^8 dpm μg^{-1} DNA, Dekaprime kit, Fermentas, http://www.thermoscientificbio.com/fermentas/). The membrane was washed with both 2× SSC, 0.1% w/v SDS (2× 5 min) and 0.2× SSC, 0.1% w/v SDS (15 min each). The membrane was imaged using a Storm phosphorimager (Molecular Dynamics, http://www.gelifesciences.com/). The membrane was subsequently reprobed using α–^{32}P-dCTP-labelled PCR products from primer pair 3 as above.

ACKNOWLEDGEMENTS

We thank Richard Nichols and Steven Dodsworth (both Queen Mary University of London, School of Biological and Chemical Sciences) as well as Mike Fay (Royal Botanic Gardens Kew, Jodrell Laboratory) for helpful comments. We thank the Leonardo da Vinci programme for funding H.B. and the Marie Curie programme for funding L.M.; this work was also supported by the Natural Environment Research Council (grant number NE/G01724/1) and the Czech Science Foundation (P501/13/10057S). This paper includes Illumina data generated by the Centre of Genomic Research, which is based at the University of Liverpool, UK. We thank Mr Laurence Hill (Richmond, Surrey) for plant material.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Dot plot of the consensus sequences of loops 1–3 before trimming.

Figure S2. Phylogenetic placement of FriEPRV.

Figure S3. Length distribution of small RNAs matching the FriEPRV consensus sequences.

Figure S4. Graph showing the number of SNPs observed relative to the consensus sequence of loop 1.

Table S1. Primers and PCR conditions used.

Table S2. Plant material studied.

Data S1. Contigs and consensus sequences.

Appendix S1. Estimation of the FriEPRV genome proportion (GP).
Appendix S2. Copy number of FriEPRV.
Appendix S3. Cytosine methylation analysis: SNPs in genomic reads.

REFERENCES

Bennett, M.D., Leitch, I.J., Price, H.J. and Johnston, J.S. (2003) Comparisons with *Caenorhabditis* (~100 Mb) and *Drosophila* (~175 Mb) using flow cytometry show genome size in *Arabidopsis* to be ~157 Mb and thus 25% larger than the *Arabidopsis* Genome Initiative estimate of ~125 Mb. *Ann. Bot.* **91**, 547–557.

Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580.

Chabannes, M. and Iskra-Caruana, M.-L. (2013) Endogenous pararetroviruses - a reservoir of virus infection in plants. *Curr. Opin. Virol.* **3**, 615–620.

Chabannes, M., Baurens, F.C., Duroy, P.O., Bocs, S., Vernerey, M.S., Rodier-Goud, M., Barbe, V., Gayral, P. and Iskra-Caruana, M.L. (2013) Three infectious viral species lying in wait in the banana genome. *J. Virol.* **87**, 8624–8637.

Day, P.D., Berger, M., Hill, L., Fay, M.F., Leitch, A.R., Leitch, I.J. and Kelly, L.J. (2014) Evolutionary relationships in the medicinally important genus *Fritillaria* L. (Liliaceae). *Mol. Phylogen. Evol.* **80**, 11–19.

Di Tommaso, P., Moretti, S., Xenarios, I., Orobitg, M., Montanyola, A., Chang, J.-M., Taly, J.-F. and Notredame, C. (2011) T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Res.* **39**, W13–W17.

Grover, C. and Wendel, J.F. (2010) Recent insights into mechanisms of genome size change in plants. *J. Bot.* **2010**, Article ID 382732.

Hansen, C.N., Harper, G. and Heslop-Harrison, J.S. (2005) Characterisation of pararetrovirus-like sequences in the genome of potato (*Solanum tuberosum*). *Cytogenet. Genome Res.* **110**, 559–565.

Hawkins, J.S., Proulx, S.R., Rapp, R.A. and Wendel, J.F. (2009) Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. *Proc. Natl Acad. Sci. USA*, **106**, 17811–17816.

Hu, T.T., Pattyn, P., Bakker, E.G. *et al.* (2011) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat. Genet.* **43**, 476–481.

Jakowitsch, J., Mette, M.F., van der Winden, J., Matzke, M.A. and Matzke, A.J.M. (1999) Integrated pararetroviral sequences define a unique class of dispersed repetitive DNA in plants. *Proc. Natl Acad. Sci. USA*, **96**, 13241–13246.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O. and Walichiewicz, J. (2005) Repbase update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467.

Kanizay, L. and Dawe, R.K. (2009) Centromeres: long intergenic spaces with adaptive features. *Funct. Integr. Genomics*, **9**, 287–292.

Kanizay, L.B., Albert, P.S., Birchler, J.A. and Dawe, R.K. (2013) Intragenomic conflict between the two major knob repeats of maize. *Genetics*, **194**, 81–89.

Kelly, L.J. and Leitch, I.J. (2011) Exploring giant plant genomes with next-generation sequencing technology. *Chromosome Res.* **19**, 939–953.

Kelly, L.J., Leitch, A.R., Fay, M.F., Renny-Byfield, S., Pellicer, J., Macas, J. and Leitch, I.J. (2012) Why size really matters when sequencing plant genomes. *Plant Ecol. Divers.* **5**, 415–425.

Kelly, L.J., Leitch, A.R., Clarkson, J.J., Knapp, S. and Chase, M.W. (2013) Reconstructing the complex evolutionary origin of wild allopolyploid tobaccos (*Nicotiana* section *Suaveolentes*). *Evolution*, **67**, 80–94.

Kim, M.Y. and Zilberman, D. (2014) DNA methylation as a system of plant genomic immunity. *Trends Plant Sci.* **19**, 320–326.

Kim, S., Park, M., Yeom, S.-I. *et al.* (2014) Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. *Nat. Genet.* **46**, 270–278.

Koukalova, B., Moraes, A.P., Renny-Byfield, S., Matyasek, R., Leitch, A.R. and Kovarik, A. (2010) Fall and rise of satellite repeats in allopolyploids of *Nicotiana* over c. 5 million years. *New Phytol.* **186**, 148–160.

Kovarík, A., Koukalová, B., Lim, K.Y., Matyásek, R., Lichtenstein, C.P., Leitch, A.R. and Bezdek, M. (2000) Comparative analysis of DNA methylation in tobacco heterochromatic sequences. *Chromosome Res.* **8**, 527–541.

Krueger, F. and Andrews, S.R. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, **27**, 1571–1572.

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25.

Leitch, A.R. and Leitch, I.J. (2012) Ecological and genetic factors linked to contrasting genome dynamics in seed plants. *New Phytol.* **194**, 629–646.

Leitch, I.J. and Leitch, A.R. (2013) Genome size diversity and evolution in land plants. In *Plant Genome Diversity*, **vol 2**, Physical structure, behaviour and evolution of plant genomes (Leitch, I.J., Greilhuber, J., Doležel, J. and Wendel, J.F., eds). Wien: Springer-Verlag, pp. 307–322.

Leitch, I.J., Beaulieu, J.M., Cheung, K., Hanson, L., Lysak, M. and Fay, M.F. (2007) Punctuated genome size evolution in Liliaceae. *J. Evol. Biol.* **20**, 2296–2308.

Lim, K.Y., Kovarik, A., Matyasek, R., Chase, M.W., Knapp, S., McCarthy, E., Clarkson, J.J. and Leitch, A.R. (2006) Comparative genomics and repetitive sequence divergence in the species of diploid *Nicotiana* section *Alatae*. *Plant J.* **48**, 907–919.

Llorens, C., Munoz-Pomer, A., Bernad, L., Botella, H. and Moya, A. (2009) Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. *Biol. Direct*, **4**, 41.

Macas, J., Neumann, P. and Navrátilová, A. (2007) Repetitive DNA in the pea (*Pisum sativum* L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and Medicago truncatula. *BMC Genomics*, **8**, 427

Malik, H.S. and Henikoff, S. (2009) Major evolutionary transitions in centromere complexity. *Cell*, **138**, 1067–1082.

Marchler-Bauer, A., Lu, S., Anderson, J.B. *et al.* (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.* **39**, D225–D229.

Matzke, M.A. and Mosher, R.A. (2014) RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nat. Rev. Genet.* **15**, 394–408.

Nei, M. and Rooney, A.P. (2005) Concerted and birth-and-death evolution of multigene families. *Annu. Rev. Genet.* **39**, 121–152.

Neumann, P., Navrátilová, A., Koblížková, A., Kejnovský, E., Hřibová, E., Hobza, R., Widmer, A., Doležel, J. and Macas, J. (2011) Plant centromeric retrotransposons: a structural and cytogenetic perspective. *Mobile DNA*, **2**, 4.

Noreen, F., Akbergenov, R., Hohn, T. and Richert-Poggeler, K.R. (2007) Distinct expression of endogenous Petunia vein clearing virus and the DNA transposon dTph1 in two *Petunia hybrida* lines is correlated with differences in histone modification and siRNA production. *Plant J.* **50**, 219–229.

Notredame, C., Higgins, D.G. and Heringa, J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217.

Novák, P., Neumann, P. and Macas, J. (2010) Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics*, **11**, 378.

Novák, P., Neumann, P., Pech, J., Steinhaisl, J. and Macas, J. (2013) RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, **29**, 792–793.

Nystedt, B., Street, N.R., Wetterbom, A. *et al.* (2013) The Norway spruce genome sequence and conifer genome evolution. *Nature*, **497**, 579–584.

Peng, J.C. and Karpen, G.H. (2008) Epigenetic regulation of heterochromatic DNA stability. *Curr. Opin. Genet. Dev.* **18**, 204–211.

Renny-Byfield, S., Chester, M., Kovařík, A. *et al.* (2011) Next generation sequencing reveals genome downsizing in allotetraploid *Nicotiana tabacum*, predominantly through the elimination of paternally derived repetitive DNAs. *Mol. Biol. Evol.* **28**, 2843–2854.

Renny-Byfield, S., Kovarik, A., Kelly, L.J., Macas, J., Novak, P., Chase, M.W., Nichols, R.A., Pancholi, M.R., Grandbastien, M.A. and Leitch, A.R. (2013) Diploidization and genome size change in allopolyploids is associated with differential dynamics of low- and high-copy sequences. *Plant J.* **74**, 829–839.

Richert-Pöggeler, K.R. and Shepherd, R.J. (1997) Petunia vein-clearing virus: a plant pararetrovirus with the core sequences for an integrase function. *Virology*, **236**, 137–146.

Richert-Pöggeler, K.R., Noreen, F., Schwarzacher, T., Harper, G. and Hohn, T. (2003) Induction of infectious petunia vein clearing (pararetro) virus from endogenous provirus in petunia. *EMBO J.* **22**, 4836–4845.

Simon, S.A. and Meyers, B.C. (2011) Small RNA-mediated epigenetic modifications in plants. *Curr. Opin. Plant Biol.* **14**, 148–155.

Staginnus, C., Gregor, W., Mette, M.F., Teo, C., Borroto-Fernandez, E., Machado, M.L., Matzke, M. and Schwarzacher, T. (2007) Endogenous pararetroviral sequences in tomato (*Solanum lycopersicum*) and related species. *BMC Plant Biol.* **7**, 24.

Swofford, D.L. (2003) *PAUP*: Phylogenetic Analysis Using Parsimony (*and other methods)* Version 4. Sunderland, MA: Sinauer Associates.

Teycheney, P.Y. and Geering, A. (2011) Endogenous viral sequences in plant genomes. In *Recent Advances in Plant Virology* (Caranta, C., Aranda, M.A., Tepfer, M., Lopez-Moya, J.J. and Caister, eds). Norfolk: Academic Press, pp. 343–362.

Zemach, A., Kim, M.Y., Hsieh, P.H., Coleman-Derr, D., Eshed-Williams, L., Thao, K., Harmer, S.L. and Zilberman, D. (2013) The Arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. *Cell*, **153**, 193–205.

Zhang, H., Koblízková, A., Wang, K. *et al.* (2014) Boom-bust turnovers of megabase-sized centromeric DNA in *Solanum* species: rapid evolution of DNA sequences associated with centromeres. *Plant Cell*, **26**, 1436–1447.